A dynamic model of deciding not to choose

Angus Reynolds[1], Roderick Garton[1], Peter Kvam[2], James Sauer[1], Adam Osth[3]
and
Andrew Heathcote[1].

[1]Department of Psychology, University of Tasmania
[2] Department of Psychology, University of Florida, Formerly Ohio State University
[3] Melbourne School of Psychological Sciences, University of Melbourne.

We propose a dynamic theory of decisions not to choose which of two options is correct. Such "don't-know" judgements are of theoretical and practical importance in domains ranging from comparative psychology, psychophysics, episodic memory and metacognition to applied areas including educational testing and eyewitness testimony. However, no previous theory has provided a detailed quantitative account of the time it takes to make both definitive and don't-know responses and their relative frequencies. We tested our theory, the "Multiple Threshold Race" (MTR), in one recognition memory experiment where participants had to pick a previously studied target out of two similar faces and another where targets and lures were tested one at a time. In both experiments we manipulated similarity through face morphing. High similarity made decisions difficult, encouraging don't-know responses. We also tested the MTR's ability to account for other manipulations that aimed to affect the speed and probability of don't-know responses, including increasing penalties for making an error (with no penalty for a don't-know response) and emphasising either response speed or accuracy. We found that there were marked individual differences in don't-know use, and that the MTR was able to account for the intricate pattern of effects associated with our manipulations, both on average and in terms of individual differences. We discuss how estimates of MTR's parameters illuminate the psychological mechanisms that govern the interplay between definitive and don't-know responding.

Introduction

Sometimes it is important to know when you don't know what to choose. For example, when faced with a decision based on uncertain information it may be more prudent to put off making a definitive choice, and instead collect further information (Busemeyer & Rapoport, 1988). In a scenario familiar in game shows where there is a fixed period of time to give answers, it can be advantageous to quickly pass on questions to which you think you don't know the answer in order to move on to the next question. The scenario that we focus on here occurs with a fixed number of uncertain choices between two options where there are losses for errors and gains for being correct. As long as it is possible to discern choices where accuracy is likely to be low, it can be worthwhile in terms of long-run returns to avoid a definitive choice if that also avoids the loss associated with an error. This sort of scenario is important in the many real-world situations, ranging from multiple-choice exams with formula scoring (where there is a penalty for wrong responses, Higham, 2007), to high-stakes decisions in eyewitness line-ups where a false identification can have dire consequences for an innocent suspect (Brewer & Wells, 2011).

The textbook advice in the early psychophysical measurement literature (e.g., Woodworth, 1938) was that such “equivocal” choices should not be allowed, both because they are made in just the cases where definitive choices are most informative (e.g., Brown, 1910; Boring, 1921; Jastrow, 1888), and in order to avoid ambiguities in measurements of sensitivity caused by large variations in the frequency with which individual participants make use of them (e.g., Angell, 1907; Boring, 1920; Fernberger, 1930). The same high degree of variability has also been noted in educational testing when a “don't-know” option was provided in multiple-choice tests, with a sizeable proportion of participants never using it (Friedman & Fleishman, 1956). The rate of don't-know use can also depend heavily on

prompts; Weber and Perfect (2012) found that in an eyewitness lineup task a spontaneous rate of ~2% rate increased ten-fold when an explicit don't-know option was provided.

In contrast to this early skepticism, Watson, Kellogg, Kawanishi and Lucas (1973) concluded that "uncertain" is a meaningful psychophysical response and can be treated as equivalent to a middle category between binary choice alternatives in a signal-detection theory analysis. They found that individual differences could be accommodated through variations in the two thresholds demarcating the three possible responses without compromising the measurement of detection sensitivity. Subsequently it was shown that such "uncertain" or "escape" responses can be used adaptively to deal with difficult choices not only by humans, but also by bottlenose dolphins (*Tursiops truncates*; Smith et al., 1995) and Rhesus monkeys (*Macaca mulatta*; Shields, Smith & Washburn, 1997). More recently, Kiani and Shadlen (2009) found that in Rhesus monkeys the same neurons in the parietal cortex that mediated choices about motion direction also represented the decision to opt out of making a choice.

In the present work we study don't-know responses in recognition-memory tasks. Perhaps reflecting early concerns in the psychophysical literature, don't-know responses have not been heavily studied in recognition memory (e.g., Clark, Howell & Davey, 2008, report their use in only 13 of 94 studies in their meta-analysis of eyewitness identification), but more recently their utility has been increasingly acknowledged. For example, in a task requiring eyewitness to either identify a suspect or reject the entire line-up, Weber and Perfect (2012) found that also providing a don't-know option made definitive responses more accurate, more diagnostic of the suspect's guilt or innocence, and did not decrease the quantity of correct decisions. At a theoretical level, Perfect and Weber (2012) extended Koriat and Goldsmith's (1996) influential model of memory monitoring, to model the

frequency of identify/don't-know/reject lineup decisions using dual-threshold signal-detection framework similar to similar to Watson et al. (1973).

In this paper we propose a new dynamic theory of “don’t-know” responses that addresses the speed with which they are made as well as their frequency. We test the theory’s ability to cope with the individual differences and the potential ambiguity of don’t-know decisions that perturbed the early psychophysicists and its ability to account for several experimental manipulations that affect the frequency and speed of responses. Our experiments focus on recognition-memory decisions about pictures of faces. The recognition-memory decisions we examined were difficult because each lure face (i.e., a test face that was not studied) was similar to a studied (target) face. We investigated how the effect of similarity interacts with the format of the recognition test over two experiments. The first experiment used two-alternative forced-choice testing, where the recognition decision was between a target and a similar lure appearing side by side. Similarity between test alternatives presented in this format is known to induce uncertainty as reflected in confidence ratings (Dobbins, Kroll, & Liu, 1998; Heathcote, Bora, & Freeman, 2010; Horry & Brewer, 2016; Tulving, 1981), so we hypothesised that it would also affect don’t-know responses. We contrasted these results with those from a second experiment where faces were tested one at a time, and so variations in lure similarity were less obvious to participants.

In both experiments we also manipulated payoffs for losses due to error responses relative to a fixed gain for correct responses, with neither gains nor losses for don't-know responses. If errors are costlier and participants are trying to maximise gains, don’t-know use should increase with greater punishments, at least if they feel that they are able to make metacognitive judgements that provide valid information about the likely accuracy of their choices. Finally, we manipulated the urgency with which participants made recognition decisions through instructions that emphasised either the importance of speed or of accuracy.

If the calibration of metacognitive judgements about the difficulty of choices is degraded by having less time available to make a decision, then participants are likely to find it more difficult to use don't-know responses to effectively manage error costs when speed is emphasised.

The key theoretical innovation that we bring to this investigation draws on Vickers' (1979, 2001) dynamic "balance of evidence" model of confidence judgement to create a quantitative cognitive model of don't-know decisions, the Multiple-Threshold Race (MTR). Because the MTR models both the frequency of definitive vs. don't-know decisions and the time it takes to make them, it enables us to account for speed vs. accuracy tradeoffs, and to use response time (RT) as well as choice proportions to inform our understanding of the cognitive processes underlying don't-know decisions. We instantiate the MTR using Brown and Heathcote's (2008) Linear Ballistic Accumulator (LBA) model of evidence accumulation. Pleskac and Busemeyer (2010) note that an accumulator model like the LBA, in combination with multiple thresholds (see also Van Zandt & Maldonado-Molina, 2004) can avoid pitfalls that have beset prior accumulator models using the balance of evidence, such as failing to accommodate the effects of time pressure on confidence.

Our instantiation of the MTR provides a fine-grained account of performance not only in terms of the proportions of correct, error and don't-know responses, but also the distribution of RTs. Further, it supports meaningful parameter estimates because it can be used as a measurement model, in the sense that we can recover the parameters used to generate simulated data from the MTR (Heathcote, Brown & Wagenmakers, 2015). This one-to-one mapping between the data and parameters allows us to go beyond checking the descriptive adequacy of the MTR to also assess the coherence of the insights its parameter estimates provide about the psychological processes underpinning don't-know decisions. It

also allows us to investigate individual differences in don't-know use, which were quite marked, in line with the results in the psychophysical and educational-testing literatures.

In the following we first describe the MTR and the mathematical details that enable us to use Bayesian methods to fit it to data and obtain parameter estimates. We then provide an overview of the two experiments and discuss how the MTR's parameters are related to the four manipulations we investigate: test format, target-lure similarity, error cost and speed/accuracy emphasis. Finally, we report the empirical results for each experiment, test the ability of the MTR to describe these results, and investigate the way in which its estimated parameters explain both the average effects of the experimental manipulations and individual differences.

The Multiple-Threshold Race

Standard evidence-accumulation architectures like the LBA map one accumulator to each response option, where each accumulator has an associated evidence total that increases over time. Each accumulator also has an evidence threshold with magnitude b that may differ between accumulators to accommodate response bias. The first accumulator total that reaches threshold triggers its corresponding response. RT equals the threshold crossing time plus the time for non-decision processes (i.e., encoding the choice stimulus and producing a motor response). In past applications, one LBA accumulator has been mapped to each possible response (Brown & Heathcote, 2008). Thus, it might seem that the simplest way to model binary choice with an additional don't-know response is with three accumulators. However, this approach does not intrinsically capture the fact that, in terms of certainty about decision, don't-know responses fall on a bivalent continuum between each of the definitive binary responses (Watson et al., 1973).

Instead, in the MTR there are only two accumulators, one for each definitive choice, but to both we add a second threshold below the first. Vickers (1979) proposed that confidence is an increasing function of the magnitude of the difference in the evidence between winning and losing accumulators (see also Van Zandt, 2000; Merkle & Van Zandt, 2006), but later noted that there was a "need to take account of the way in which such magnitudes may be converted into overt confidence ratings" (Vickers, 2001, p.153)[1]. Here we use multiple thresholds to convert the balance of evidence into another type of rating inversely associated with the confidence, a don't-know response. The idea of multiple thresholds in accumulator models has precedent in the work of Van Zandt and Maldonado-Molina (2004), where it provided a mechanism for making two responses in a sequence.

We also add a rule that determines the response contingent on the state of the evidence totals in *both* accumulators. We assume that, as in the standard architecture and Vicker's (1979) balance of evidence model, a response is made at a time determined by the first accumulator to cross the upper threshold. However, the response chosen only corresponds to the definitive choice associated with the winning accumulator if the total in the losing accumulator is still below its lower threshold. If the losing accumulator is above its lower threshold, then a don't-know response is made. Intuitively the latter case is more uncertain because there is less difference between the evidence totals at the moment of choice than when one of the definitive choices is selected, as is illustrated in Figure 1. By basing

[1] Merkle and Van Zandt (2005) proposed this was done by first calculating a relative balance of evidence score on the 0-1 interval (i.e., winning evidence divided by the sum of winning and losing evidence) then directly accessing this value to make a mapping to confidence (e.g., the probability of a confidence rating as a percentage between 5-15 is the probability of the relative score falling between 0.05 and 0.15). Our proposal differs in that the mapping is not direct, but rather is mediated by threshold placement, which gives it greater flexibility. Our approach is also consistent with the assumption made by most evidence-accumulation models that the only way to access the state of an accumulator is through threshold-crossing events.

decisions on the difference in evidence between the two accumulators – Vickers' (1979) "balance of evidence" – the MTR captures the intrinsic positioning of don't-know responses between the definitive responses.

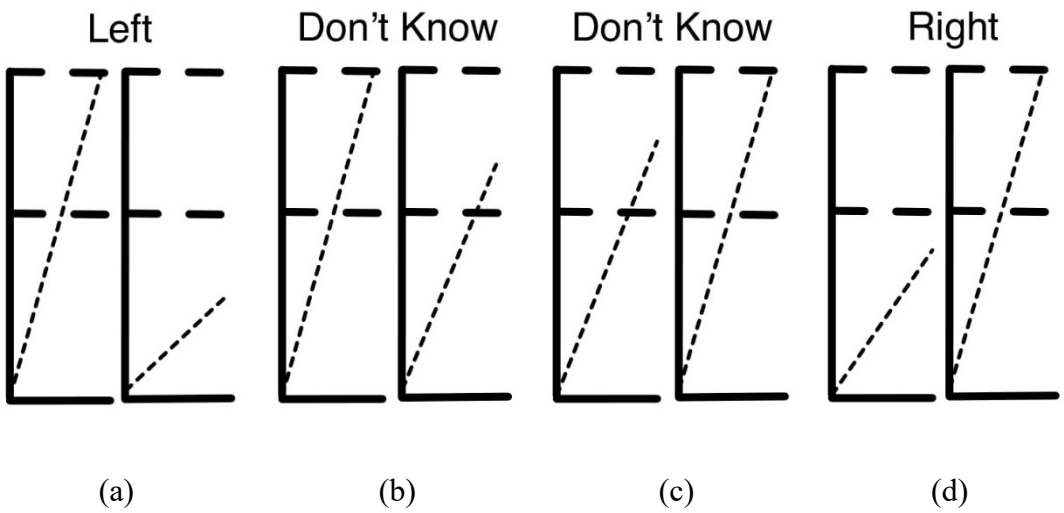


Figure 1. The for latent states of the MTR don't-know model for a binary left vs. right button-press choice. Each of (a)-(d) represent a single decision made by a pair of accumulators (which decision is made is indicated above each pair), with time on the x-axis and the evidence total on the y-axis (axes are thick solid lines). Linearly increasing evidence totals are represented by slanting dashed lines. Horizontal dashed lines represent thresholds. Each accumulator pair is represented at the moment a decision occurs and labelled above with the corresponding choice.

This choice rule means that at the unobserved or "latent" level there are two types of don't-know responses, corresponding to the accumulator mapped to one or other option winning (i.e., Figures 1b and 1c). However, at the observed or "manifest" level these two types cannot be differentiated, whereas the definitive choices remain identifiable. This ambiguity around a don't-know choice might potentially cause measurement issues, a question we return to below. We denote the magnitudes of the lower thresholds as d. As d becomes smaller the likelihood of a don't-know response increases because the range of values of the loosing accumulator corresponding to a don't-know response also increases. Furthermore, like the upper thresholds, d may vary between accumulators, modulating the proportion of don't-know responses generated by each latent state. For example, in Figure 1,

if *d* were smaller for the left accumulator than for the right accumulator then, all other things being equal, don't-know responses are more likely to be generated when the right accumulator wins than when the left accumulator wins. As *d* approaches zero for a given accumulator any case in which that accumulator loses must produce a don't-know response.

In reporting our results we use a measure defined as $DK = (1\text{-}d/b)$, which is the proportion of the region below an accumulator's choice threshold that corresponds to a don't-know response (i.e., the region between the don't-know threshold, *d*, and the choice threshold, *b*). When $DK = 1$ only don't-know responses can be made, whereas when $DK = 0$, no don't-know responses can be made. Between these values don't-know probability monotonically increases with DK.

The MTR model illustrated in Figure 1 shares many assumptions with Brown and Heathcote's (2008) LBA model, which has been shown to give a strong account of data from episodic memory paradigms (see Osth, Bora, Dennis, & Heathcote, 2017; Osth & Farrell, 2019; Rae et al., 2014). One shared assumption is that evidence totals increase linearly during a trial. Other possibilities could exist within the general MTR framework, such as racing diffusion processes (i.e., processes whose rate varies randomly during each decision). The linear deterministic assumption is useful because it makes the MTR model very computationally tractable.

Like the LBA, we assume that the rate varies randomly between trials, following a normal distribution, and we assume that distribution is truncated to only include positive values so that all accumulators have finite finishing times (see Heathcote & Love, 2012). Again, like the LBA, we also assume that the starting-point of evidence accumulation varies randomly and independently for each accumulator following a uniform distribution on the interval 0-*A*, where *A* may vary between accumulators. Start-point variability accounts for

random trial-to-trial response biases, and in terms of Figure 1, would mean that the slanting lines usually start at different points from each other.

In summary, the MTR model has six types of parameters. Five of these are shared with the LBA: non-decision time, t_0, the mean rate of evidence accumulation, v, and its standard deviation, sv, start-point noise, A, and the upper threshold, b. The MTR adds one extra parameter, d, corresponding to a second threshold. In principle all six parameters can differ between accumulators, but here we assume this is not the case for t_0 and for A without any apparent deleterious effects in fitting our data; this may differ in other applications.

In the fits reported here we enforce the orderings $0 < A < b$, by estimating parameters $A > 0$, and $B > 0$, where $B = b - A$. The ordering $A < b$ ensures that a decision cannot be made instantaneously. To ensure the definitional characteristic of the don't-know threshold, that it lies between zero and the upper or "choice" threshold, we also enforce the ordering $0 < d < b$ by estimating a parameter D on the unit interval (i.e., $0 < D < 1$) where $D = d/b$. This D parameter is then transformed using the logistic function so that it could be estimated on the real number line. Note that this definition allows that an accumulator may sometimes start in a state that can only support a don't-know response if it loses the race. We examine parameter estimates to examine the degree to which this occurs. Although we sample the value D, in analysis we are often more interested in $1\text{-}D{=}DK$, which gives the proportion of the accumulator that results in a don't know response if accumulation finishes in this region.

We define differences in rate parameters in terms of the accumulator that matches versus mismatches the stimulus. In our first experiment the accumulator corresponding to the side on which the target is presented has the matching rate, and the accumulator for the lure side has the mismatching rate. In our second experiment, if the stimulus is a target the accumulator associated with a target response has the matching rate, and the accumulator associated with a lure response has the mismatching rate, and vice versa for a lure stimulus.

Because a winning matching accumulator produces correct responses, when accuracy is greater than chance its mean rate, v_c, tends to be greater than the mean rate, v_e, for the mismatching accumulator that produces errors. We also parameterized the rate standard deviation in terms of matching and mismatching accumulators, sv_c and sv_e. In at least one experimental condition we give sv_e a fixed value of one in order to make the model identifiable (Donkin, Brown & Heathcote, 2009).

In Appendix A we provide the equations for the likelihoods of each possible response. These equations enable us to fit the MTR in a way that simultaneously takes account of all aspects of the data under the assumption that the joint distribution of responses and RTs is independently and identically distributed within participants and conditions that share the same parameters. We use hierarchical Bayesian methods that allow us to take account of multiple sources of uncertainty in our parameter estimates and garner constraint from commonalities among participants (Shiffrin, Lee, Kim & Wagenmakers, 2008).

Experiment and Model Overview

Both recognition-memory experiments required difficult choices, because it is unlikely that participants would make don't-know responses with easy choices. The two experiments differed only in the way test lists were constructed and presented, either testing recognition using pairs of items (Experiment 1) or using a single item at a time (Experiment 2). In Experiment 1 participants were tested with a horizontally arrayed pair of colour pictures of faces. If they decided to make a definitive response they had to indicate if the target was on the left or right. In Experiment 2 the definitive responses classified the single test items as a target or as a lure. In both experiments, participants studied lists of pictures of faces presented one at a time. The items in the study list were made up of one of the members of pairs of faces like those in Figure 2b or 2c. These face pairs were created by morphing

together pairs of face images chosen to be structurally similar, like the pair shown in Figure 2a. Note that these original images were never seen by participants. Pairs which were easier to discriminate were created by a 90%/10% mixture (e.g., Figure 2b) and pairs that were harder to discriminate were created by a 70%/30% mixture. This difference in the levels of morphing was chosen through a pilot experiment which confirmed that it was sufficient to produce substantially more errors and don't-know responses in single item recognition of higher vs. lower similarity lures. Half of the study items came from easy pairs and half from hard pairs. For each participant, lower and higher similarity lure items always came from a different original pair of faces.

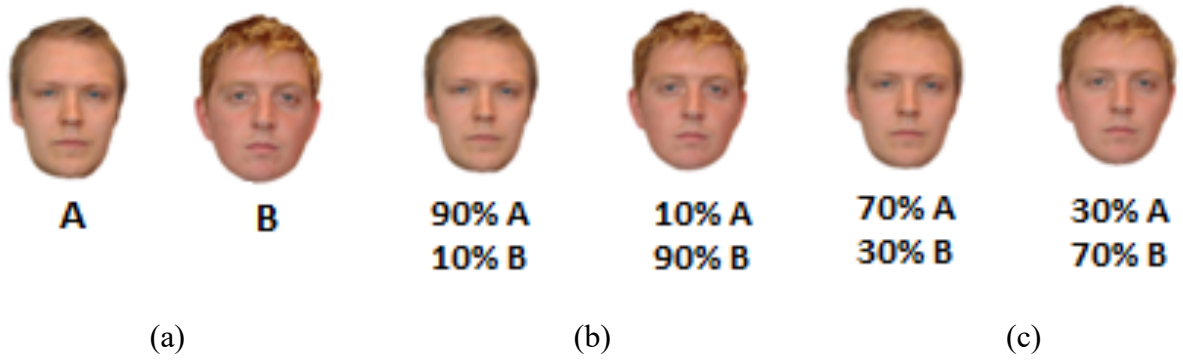


Figure 2. (a) Example of paired faces used to construct study and test images **with the same resolution as in the experiment.** (b) a corresponding lower-similarity face pair created by morphing 10% of one face with 90% of the other, and (c) a corresponding higher-similarity face pair created by morphing 30% of one face with 70% of the other.

In Experiment 1 half of the study items were drawn from lower-similarity pairs and half from higher-similarity pairs. All study items were tested, half with the target on the left and half with the target on the right, with the lure being the other member of the pair that was not studied. In Experiment 2 the study list was constructed in the same way, and half of its members were randomly selected to be test targets with equal numbers drawn from lower- and higher-similarity pairs. The other half of the test faces were lures drawn from pairs where the other member had been studied but was not tested, again with equal numbers from lower- and higher-similarity pairs. As a result, one quarter of the test trials used lower-similarity

lures and one quarter higher-similarity lures. In summary, Experiment 1 had a test stimulus factor with two levels corresponding to lower-similarity vs. higher-similarity pairs, that occurred equally often. In Experiment 2 the stimulus factor had lower-similarity lure, higher-similarity lure and target levels, occurring, respectively, on 25%, 25% and 50% of test trials.

This design ensured that face identity is not confounded with lower- vs. higher-similarity lures (i.e., there were no systematic item differences) because the same originals were drawn on to generate both hard and easy pairs. To avoid priming due to repeated presentation of items derived from the same original faces only one of the hard or easy pairs derived from the same originals were seen by a given participant. In Experiment 2 the targets that made up the remaining half of the test list were homogenous in terms of similarity because faces drawn from lower- and higher-similarity pairs were no more or less similar to other studied faces.

Both experiments included two factors that attempted to manipulate don't-know thresholds, one between subjects, *error cost*, and one within subjects, *speed vs. accuracy emphasis*, that are described in detail in the next two sections. A third within-subjects factor that attempted to manipulate choice difficulty, the *similarity of lures to targets*, differed in how it was instantiated in each experiment, and is described in the third section below. We hypothesised participants would be less likely to make don't-know responses for easier choices, reflecting their greater certainty. This type of difference in don't-know probability provides a stringent test of the model because, as we argue below and instantiate in our modelling, it must be explained without any difference in don't-know thresholds.

Error Cost

Participants might use information available before a trial to adjust the don't-know threshold, and hence the proportion of don't-know responses, in a way that can help to

achieve performance goals. In an attempt to manipulate such goals, we provided participants in both experiments with feedback based on a scoring system much like that used for formula-scored multiple-choice tests. Error cost was manipulated between subjects in order to avoid carry-over effects and so maximize any observed differences. In all cases correct responses received +100 points and no points were awarded or lost for a don't-know response.

The high-error-cost condition encouraged the use of don't-know responses by penalizing errors (-300 points) more heavily than it rewarded correct responses. In this case the optimal strategy is to respond don't-know whenever a response had less than a 75% expected chance of being correct. This is because, at 75% correct, the expected return is zero (i.e., $0.75\times100+0.25\times-300=0$); for higher accuracies the return is positive, making it better to make a definitive response; for lower accuracies the return is negative, making it better to respond don't know to obtain a sure zero return.

The low-error-cost condition penalized errors less than the high-error-cost condition (-100 points). In this case it is optimal to use a don't-know response when expected accuracy is less than 50%, as at that level the expected return is zero (i.e., $0.50\times100+0.50\times-100=0$). This is very unlikely, as 50% represents chance performance in a binary choice task. Although less than chance responding is possible for high-similarity lures, it would be circular to assume that participants can set thresholds differently for lure and target stimuli to take advantage of this occurring. Additionally, even if one could recognise that a particularly definitive response has less than 50% accuracy, it would be better to swap that response for the alternative definitive response than to make a don't-know response.

Hence, if participants are entirely governed by optimizing points, they have no reason to ever use don't-know responses in the low error-cost condition, and so should set the don't-

know threshold equal to the upper threshold (i.e., $d = b$ and so $DK = 0$). In the high error-cost condition, they should set the don't-know threshold at a lower level (i.e., $d < b$ and so $DK > 0$).

However, it is possible that participants may, at least in part, set their don't-know threshold in order to optimize an alternative criterion. For example, if they wanted to increase the accuracy of their definitive responses, they would lower their don't-know threshold, and so both increase the probability of making a don't know response and the probability that definitive responses are accurate, at least to the degree that the balance of evidence provided them with valid information about accuracy. This would encourage increased don't-know responding even in the low error-cost condition. Also, they could set the criterion so as to maximize the expected *utility* of their responses rather than the expected value. For example, if utility reflected a loss aversion, the magnitude of the utility increase for a gain of +100 points could be less than the disutility of a loss of -100 points, again encouraging don't-know responses in the low error-cost condition.

In Experiment 2 it might also be advantageous to have different don't-know thresholds for each accumulator, as target-stimulus trials will typically have higher quality evidence (i.e., a larger difference between the matching and mismatching accumulator rates) than lure trials, due to the inclusion of lures that are highly similar to targets. Hence, relative to lure trials, target trials are likely to have more correct responses, and the lure accumulator will, on average, have less evidence at the time of a correct choice. This means that a participant can afford to set a lower don't-know threshold for the lure accumulator, so that trials when the stimulus is a target and a target response is made are unlikely to trigger a don't-know response. Conversely, when the stimulus is a lure with a lower evidence quality, false target responses are more likely to become don't-know responses if the lure accumulator's don't-know threshold is lower. As the advantage of replacing wrong definitive

responses with don't-know responses varies with error cost, we allowed for asymmetries in don't-know thresholds that vary with error cost.

In Experiment 1 choices do not directly correspond to whether a stimulus is a lure or target, but rather whether it is more likely the target is on the left or right of the test pair. Hence, there is no basis to set don't know thresholds differently for each accumulator. However, it is still possible that participants have a capricious preference for one side or the other, so we accommodated that possibility by also allowing different don't-know thresholds for each accumulator that again could vary with error cost. As is conventional in evidence-accumulation modelling, preferences for either response (i.e., left vs. right in Experiment 1 and target vs. lure in Experiment 2) were also accommodated by allowing different choice thresholds for each accumulator, and again we allowed these to vary over error-cost conditions.

Speed vs. Accuracy Emphasis

In both experiments we manipulated instructions that emphasised either the speed or accuracy of responding. These instructions were manipulated in a blocked manner (i.e., between lists) in order to make it easier for participants to follow them. We used a speed-accuracy manipulation to study how it affects the frequency and speed of don't-know responses. As accuracy emphasis instructions pertain to definitive responses, participants may be more likely to try to use don't-know responses to improve definitive-response accuracy under accuracy emphasis. As previously discussed, this may occur even in the low error-cost condition that does not reward making don't-know responses. Effects on response frequency may also arise if definitive responses and don't-know responses differ in speed; if don't-know responses are faster, they may be favoured under speed emphasis; if they are slower, they may be favoured under accuracy emphasis. In light of these considerations we

allowed choice and don't-know thresholds to vary freely between speed and accuracy conditions.

The speed vs. accuracy manipulation produces a benchmark effect on speed for evidence-accumulation models of binary choice: errors that are slower than correct responses under accuracy emphasis and errors that are as fast or faster than correct responses under speed emphasis (e.g., Ratcliff & Rouder, 1998). It was originally thought that speed-accuracy emphasis selectively influenced thresholds, reducing them under speed emphasis. Reduced thresholds both speeds responding (because it takes less time to accumulate enough evidence to reach a lower threshold) and reduces accuracy (because shorter accumulation increases noise from factors like random biases present at the start of accumulation). However, there is now a consensus that rates can also be affected (e.g., Rae, Heathcote, Donkin & Brown, 2014), at least when the speed-accuracy manipulation is sufficiently potent (Starns, Ratcliff & McKoon, 2012). In accumulator models like the LBA the rate effect generally takes the form of speed emphasis increasing both the match and mismatch rates, which speeds responding, but decreasing the difference between them, which decreases accuracy. Given these results, in our fits of the MTR model we allowed mean rates to vary with speed-accuracy emphasis in both experiments.

Similarity and Test Format.

Previous research indicates that when two similar recognition test items are presented side by side, participants can, at least to some degree, discount common features, protecting them, at least to some degree, against the reduction in accuracy caused by interference due to similarity (Wixted & Mickes, 2014). Hence, it is likely that the effect of the similarity manipulation will be weaker in Experiment 1, where target and lure are presented side by side, than in Experiment 2, where only one test item is presented at a time and so there is no

clear way for participants to identify which are the best features to discount. In the MTR, as in the LBA, accuracy is largely determined by the size of the advantage of the rate of the matching (v_c) over the mismatching (v_e) accumulator. Hence it is likely that this rate difference (i.e., v_c–v_e) between high and low similarity conditions will be attenuated in Experiment 1 relative to the difference in v_c–v_e for high vs. low similarity lures in Experiment 2. Such differences in rates are allowed as we fit each experiment separately. In Experiment 1 we freely estimated separate matching and mismatching rates for high and low similarity pairs. In Experiment 2 we freely estimated separate matching and mismatching rates for targets and high and low similarity lures.

Because they so strongly resemble targets, accuracy for high similarity lures can potentially be systematically below chance, at least for some participants, and particularly in Experiment 2. The MTR can accommodate this possibility by allowing estimates of $v_c < v_e$. Even when accuracy is above chance, $v_c > v_e$ may not hold for one of the two stimuli when there is a sufficiently strong response bias for that stimulus (i.e., a lower value of b for the corresponding accumulator), because an accumulator with a slower rate can win if it has to travel less distance to reach its threshold. All of these possibilities were accommodated in Experiment 2 by estimating different values of v_c and v_e for targets, high similarity lures and low similarity lures. In Experiment 1 v_c is the rate for the accumulator corresponding to the side on which the target was presented and v_e the accumulator corresponding to the side on which the lure was presented. Different values of v_e were estimated for low and high similarity lures, and different values of v_c for targets that accompanied high and low similarity lures.

In evidence-accumulation models it is assumed that thresholds cannot be changed based on stimulus properties about which a decision is being made. To assume otherwise would be circular, as that would require knowledge of the thing which is being decided on

each trial. Hence, we make the same assumption for don't-know thresholds. This has the implication that all thresholds must be the same for targets, lower similarity lures and higher similarity lures in Experiment 2. This selective-influence assumption, which means that only rate parameters vary over the stimulus factor provides a strong test of the MTR because it must account for effects of similarity on don't-know responding without any change in don't-know thresholds.

Individual Differences

In both experiments we found very large individual differences in don't-know use, with some participants being very resistant to making don't-know responses even in the high error-cost condition. This was the case despite instructions that clearly outlined their potential benefits and demonstrations that they in fact often did not know the right answer because their recognition performance was very error prone due of the difficult nature of the choice. These individual differences present a challenge because they necessarily reduce the reliability of differences between low and high error-cost groups. However, they also represent an opportunity, because they provide both a test of the MTR model's ability to fit this variation between individuals and psychological insights into the causes of these differences in terms of relationships among MTR parameter estimates.

In order to realize this opportunity, we report not only the fit of the model to group-averaged data, but also its fit to data at the level of individual participants. Because the probability of making don't-know responses increases with the MTR's *DK* parameter for each accumulator we examined the relationship between the average value *DK* over accumulators and individual differences in making don't-know responses. However, it is important to note that other model parameters also affect the probability of making don't-

know responses, with the degree of divergence between individual *DK* estimates and don't know use quantifying their importance.

Experiment 1

Method

Participants

In total 43 subject participated. All attended a one-hour session and were University of Tasmania students with ages in years ranged from 17 to 62. Psychology students received course credit for participation, and students from other faculties was compensated for expenses in attending with a $20 shopping voucher. All participants provided prior informed consent to participate and to publicly disseminate their de-identified data. All aspects of the study relevant to participation were approved by the Tasmanian Social Sciences Human Research Ethics Committee (Ref No. H0012660)

The participants were randomly assigned to the error-cost condition, resulting in 23 with high error-cost and 20 with low error-cost. Participants were run in groups ranging from 1 to 4 in size. All participants provided prior informed consent to participate and to publicly disseminate their de-identified data.

Stimuli

The stimulus pool was comprised of digitally stored colour photographs of faces retrieved from several internet databases. The faces were of adults approximately equally sampled from both sexes, representing a variety of ages and cultures but predominantly young adults of seemingly Anglo-European backgrounds. The images included all of a person's hair (where present) and ears (where visible) but were cropped as much as possible

about the chin-line, excluding the neck and shoulders. No faces bore facial hair or glasses, and any distinguishing features (e.g., birthmarks, earrings) were removed.

In order to render two levels of recognition difficulty, 336 pairs of faces were subjectively matched on the basis of similar face shape, hairstyle and hair colour. Then, each face was digitally morphed (using FantaMorph 5 software from Abrosoft) as depicted in Figure 2. In this way, each item in the 672-item stimulus pool was paired with a face from which it could be discriminated with more or less difficulty.

Procedure

After introductory verbal and on-screen instructions, the sessions were composed of two sets of six study-test cycles, each preceded by a practice cycle, and each of which lasted about 15 minutes. These sets were differentiated only with respect to the speed or accuracy instruction that preceded them and were separated by a self-paced break. For speed-stress, participants were instructed to try to respond as quickly as possible on the basis of the first decision they could make with any degree of accuracy. For accuracy-stress, participants were instructed to use as much of the available time as necessary to ensure they made a response that was as accurate as possible. An icon was permanently displayed in the top-left corner of the screen, indicating whether the cycle required emphasis on speed or accuracy. Participants were also instructed to try to maximize their tally of points, and it was emphasised that the don't-know response option should be used in order to achieve this objective.

Introductory screen instructions preceded each cycle that reminded participants to be ready to respond, with their fingers on the appropriate response-keys. The onset of each cycle was initiated by the participant by pressing any keyboard button. The study phase of each cycle involved presentation of 28 faces. Presentation commenced with a central fixation point for 0.5s, and then each face, in the centre of the screen, for 1s, separated by a blank inter-

stimulus interval of 0.5s. The test phase commenced 1s after the last study face. Test stimuli for each cycle were constituted of 25 targets – 12 from high-similarity pairs and 12 from low-similarity pairs – paired with corresponding lures; 24 targets were drawn from the central 24 study items and one from the remainder. The middle items were selected to minimize variation due to study-order primacy and recency effects. Responses to the single item drawn from the first- and last-two studied items were not analysed for the same reason, with this item included in the test list so that participants were less likely to be aware of the exclusion of primacy and recency items. The location of targets in pairs, and the order in which pairs were tested, was randomized. Participants pressed the "z" key to indicate that the left face was the target and the "/" key to indicate the right face was the target or the space-bar key to indicate a don't-know response.

Test trials commenced with a central fixation point for 1s, followed by a test stimulus for a maximum of 2.5s, accuracy feedback for 1s, and a blank interval for 0.5s. The test item was immediately offset upon registration of a response. Accuracy feedback was in the form of centrally presented text-string that showed the amount gained (100) if correct, lost (−100 or −300) if incorrect, or the word "skipped" if the don't-know response had been used. Alternatively, a "too slow" message was presented in the case of a response timeout. After a short interval this string quickly moved up to the top-right of the screen where its value updated the tally. This tally then momentarily flickered so as to draw attention to the updated value. Any points already acquired could be preserved, but none were gained, by a don't-know response. The numerical tally of points was permanently displayed in the top-right corner of the screen during the test phase. In order to avoid loss of motivation, the tally could not decrease below zero, such that errors incurred at this point were not costly.

The stimulus faces were presented against a constant grey-and-white patterned background that differed between study and test phases so as to encourage face recognition rather than picture recognition. The background measured 15 cm × 12 cm. Responses were made using either the “z” or “/” keys on a computer keyboard for yes or no responses, with this attribution alternated in the order participants were tested, and the space bar for don't-know responses. Participants were free to use either the index or middle fingers of each hand for the “z” or “/” keys, and thumbs for the space-bar, and were instructed to keep these fingers in place upon the keyboard at all times, and so to be equally prepared to execute a response from any of the alternatives as appropriate. A graphical reminder of this mapping was permanently displayed at the bottom of the screen throughout the run.

Results

We first analysed definitive responses, both in terms of their accuracy and mean RT, then examined the probability of making don't-know response and compared don't-know mean RT to that of definitive responses. Analyses were conducted with the *lme4* R package (Bates, Maechler, Bolker, & Walker, 2014), using a Gaussian error model for the logarithm of RT and a binomial probit model for the binary choice data. ANOVA inferences were made via Wald χ^2 tests with type III sums of squares as implemented by the *car* package (Fox et al., 2012) and we judge significance at a $p < .01$ criterion.

Subsequently we report parameter estimates for MTR fit separately to the high and low-error-cost conditions. Credible intervals (e.g., ranges in which 95% of the posterior parameter estimates fall) and corresponding probabilities used for inference when comparing parameters (e.g., the probability that a difference in parameters is greater than zero) were calculated based on the proportions of posterior parameter samples averaged over participants for each MCMC iteration. The model had 19 parameters in total, with this number (being

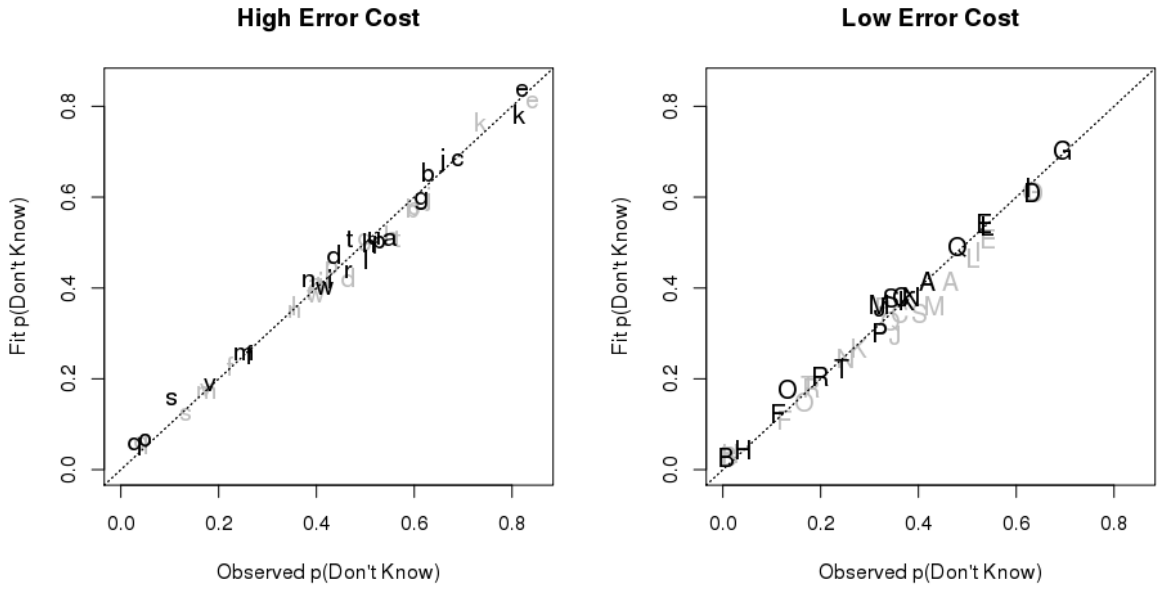
only a little more than one per accumulator for each cell of the 8 within-subject experimental conditions), being necessary to accommodate the complex design. There were four don't-know (*d*) and four choice (*b*) parameters, one for the left and one for the right accumulator separately for the speed and accuracy conditions. There was one start-point noise (*A*), one non-decision time (t_0) and one rate standard deviation (*sv*) parameter for the matching accumulator for all conditions. The *sv* parameter for the mismatching accumulator was fixed at one to make the model identifiable[2]. There were eight mean rate parameters, half for the matching accumulator and half for the mismatching accumulator, with different values for lower- and higher-similarity pairs in the speed and accuracy conditions. For details of sampling methods and priors see supplementary materials. Data files and Dynamic Models of Choice (Heathcote et al., 2018) R code to fit the models are available at osf.io/6h4qe/

Rather than report every detail of a rather complicated design, we graph the substantial effects (as confirmed by ANOVA inference) in mean RT and response probability. On each graph we also plot the corresponding effects for the fitted model. In supplementary materials we plot the global fit of the model to all conditions, both in terms of RT distribution (as represented by the 10th, 50th and 90th percentiles) and response probability; the global account provided by model is quite good in all conditions. To further illustrate that the model provides an accurate account of the full distribution of RT, observed and fitted cumulative density functions are also provided in supplementary materials.

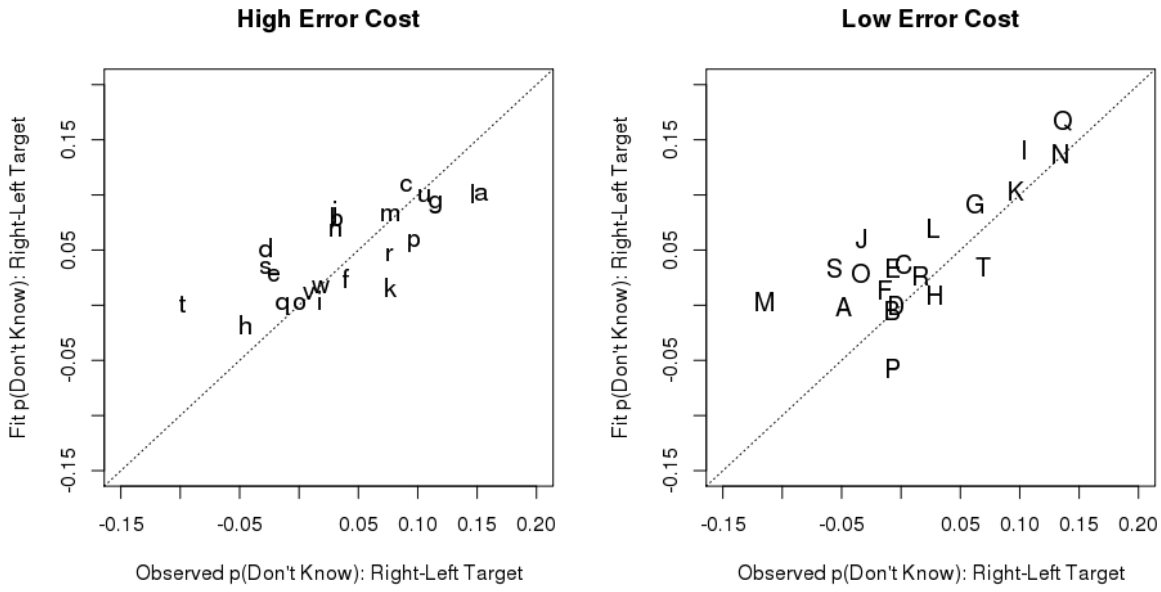
To set the scene in terms of the individual differences in don't-know use (from negligible to over 80%) and to illustrate the model's ability to accommodate them, Figures 3a

[2] Heathcote and Love (2012) showed that allowing for a difference in *sv* between matching and mismatching accumulators clearly improved the fit of the LBA and allowed it to better account for differences in the speed of correct and error responses. The same was true in the fits reported here.

and 3b plot observed against fitted don't-know probabilities for each participant separately for the high and low error-cost conditions. Comparison of the two panels shows that don't-know use was greater when error cost was higher, as expected (see also Figures 4a and 4b). There was also tendency for don't-know use to be more frequent when the target was on the right (black symbols) than on the left (grey symbols). Figures 3c and 3d plot the right-left differences, which in some cases were as much as 15%, although for a minority of participants this was reversed by as much as 10%, further underlining the marked extent of individual variation. Again, the model provides a good account of this effect, although with a few outlying participants, such as “t” for high error cost and “M” for low error cost.



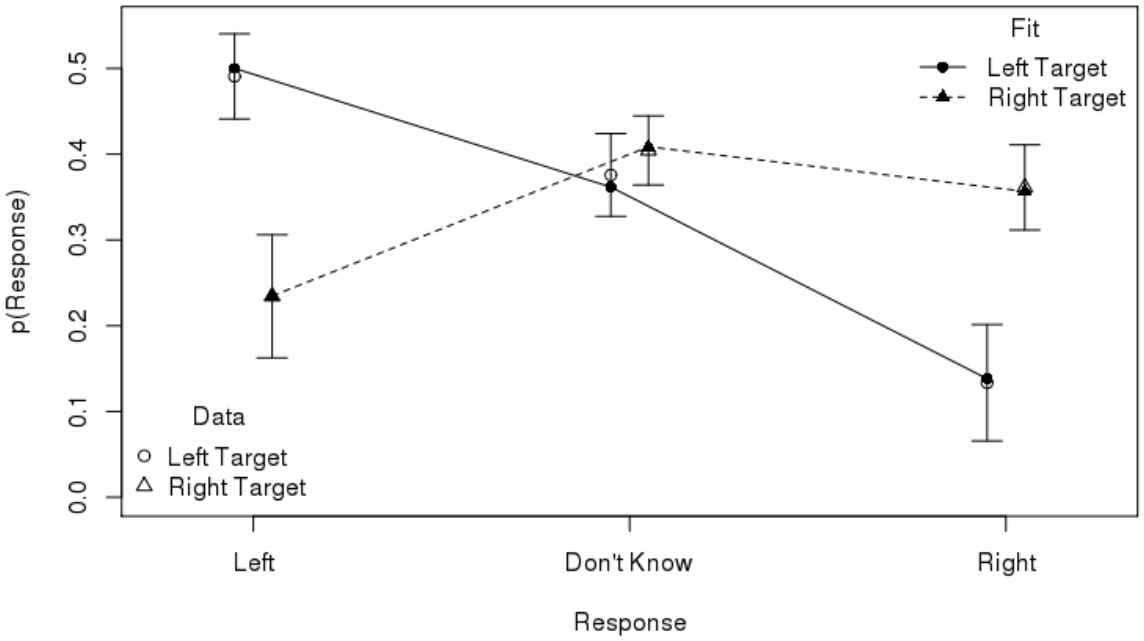
(a) (b)



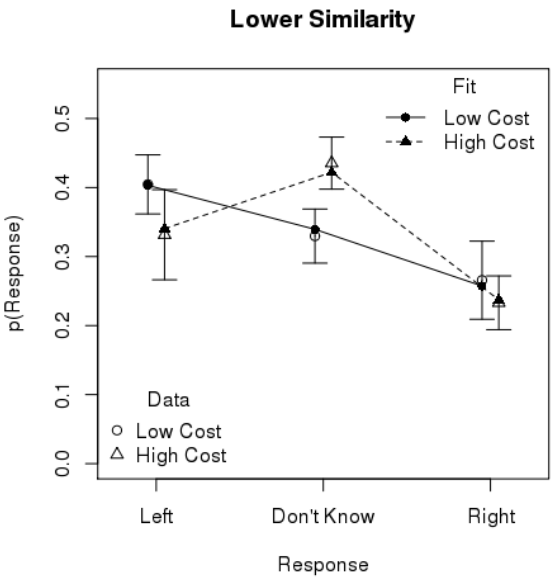
(c) (d)

Figure 3. (a) and (b): Observed vs. fitted don't-know probability for left targets (grey letters) and right targets (black letters). Letters correspond to participants: a ... w for the 23 participants in the high error-cost condition and A … T for the 20 participants in the low error cost-condition. (c) and (d) observed and fitted differences between right and left target results.

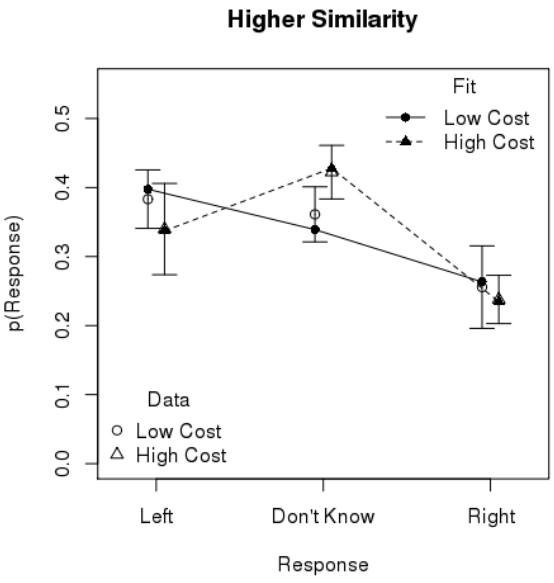
Figure 4a shows the higher average don't-know probability for right targets, $\chi^2(1) = 12.5$, $p < .001$. The only other significant effect on don't-know probability was an interaction between similarity and error cost. As shown in Figure 4b and 4c, don't-know responses were more common with higher than lower error cost, and the difference was larger for lower than higher similarity pairs, $\chi^2(1) = 7.8$, $p = .005$. Figure 4b and 4c also show an advantage in accuracy of definitive responses for left targets (78.4%) over right targets (72.8%), $\chi^2(1) = 299$, $p < .001$. Figure 4d shows the only other significant effect on the proportion of correct definitive response, a small but consistent advantage for the accuracy-stress condition (76.3%) over the speed-stress condition (75.2%), $\chi^2(1) = 12.5$, $p < .001$. It also shows that that there was no evidence for an effect of speed vs. accuracy emphasis on don't-know use.


Fit
Left Target
Right Target
p(Response)
0.0
0.1
0.2
0.3
0.4
0.5
Data
Left Target
Right Target
Left
Don't Know
Right
Response

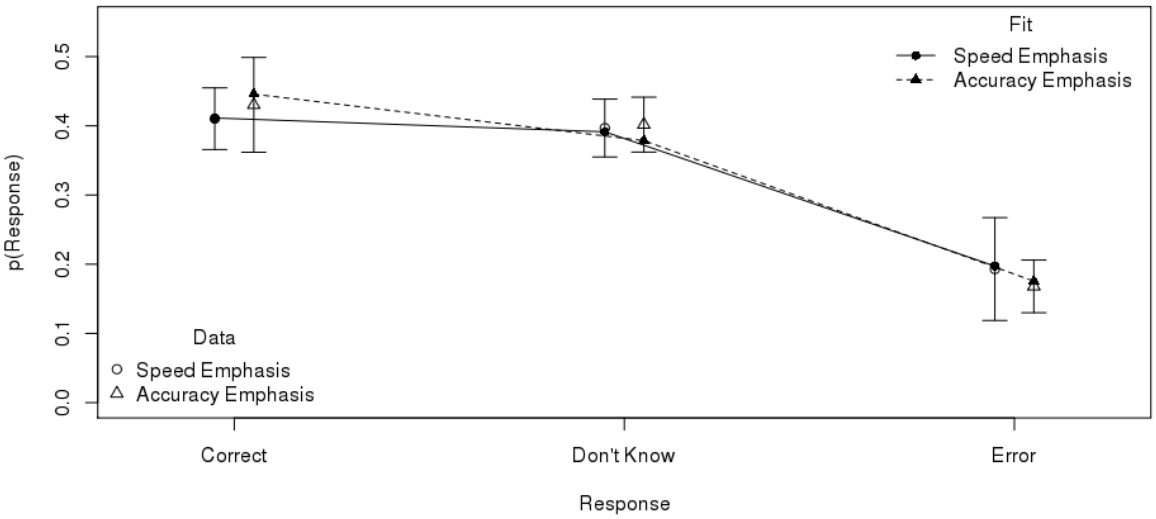
(a)


Lower Similarity
Fit
Low Cost
High Cost
p(Response)
0.0
0.1
0.2
0.3
0.4
0.5
Data
Low Cost
High Cost
Left
Don't Know
Right
Response

(b)

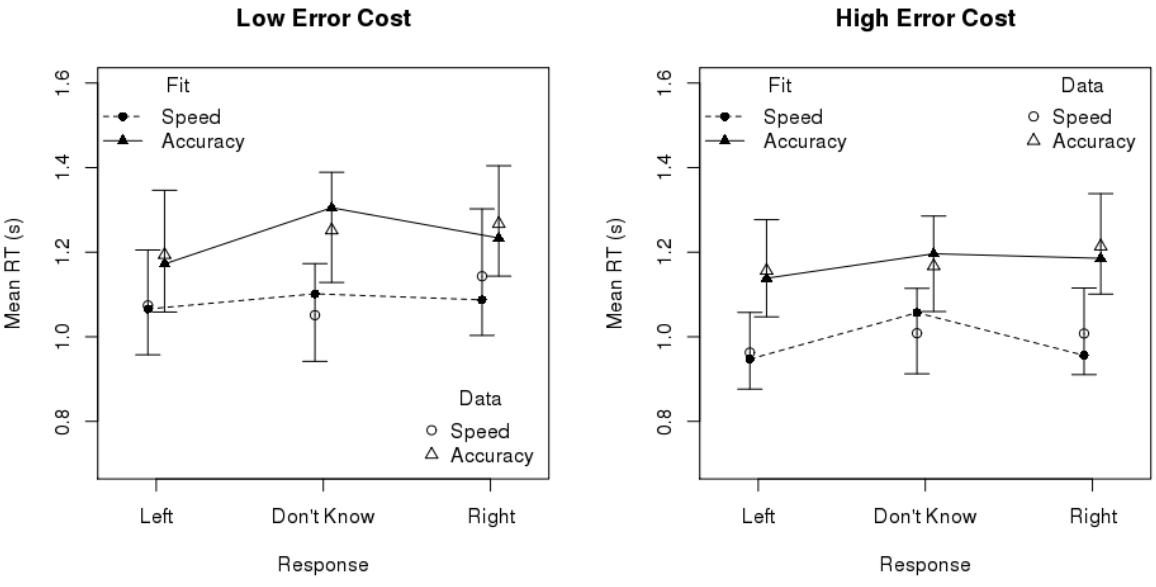

Higher Similarity
Fit
Low Cost
High Cost
p(Response)
0.0
0.1
0.2
0.3
0.4
0.5
Data
Low Cost
High Cost
Left
Don't Know
Right
Response

(c)



(d)

Figure 4. Response probability for data (open symbols with 95% confidence intervals) and fits (closed symbols joined by lines) as a function of response probability for: (a) left and right targets; lower and higher similarity pairs in (b) low and (c) high error cost and (d) speed vs. accuracy emphasis.



(a) (b)

Figure 5. Mean RT for data (open symbols with 95% confidence intervals) and fits (closed symbols joined by lines) as a function of response broken down by speed vs. accuracy emphasis in the (a) low error-cost and (b) high error-cost conditions.

Although speed vs. accuracy emphasis had only a small effect on response probability, Figure 5 shows that it had a large overall effect on RT, $\chi^2(1) = 1535$, $p < .001$, that was generally larger for high than low error cost, $\chi^2(1) = 21.4$, $p < .001$. The speed vs. accuracy effect also interacted with response, $\chi^2(1) = 30.8$, $p < .001$, with the largest difference for don't-know followed by right then left responses. As can be seen by comparing Figures 5a and 5b this interaction varied with error cost, $\chi^2(1) = 15.4$, $p < .001$, with the difference in don't-know largest for low error cost (~ .2s) but smallest for high error cost (~.16s), whereas left and right response differences were smallest for low error cost (~ .12s) and largest for high error cost (~.2s). Overall, right responses (~1.15s) were slower than left and don't-know responses (~1.1s), $\chi^2(1) = 57.6$, $p < .001$. Responses to left targets (~1.1s) were slightly faster than to right targets (~1.125s), $\chi^2(1) = 6.7$, $p < .01$, and this effect interacted and response,), $\chi^2(1) = 30.7$, $p < .001$, with the largest difference in definitive responses (~.2s) and virtually no difference for don't-know responses. No other effects were significant.

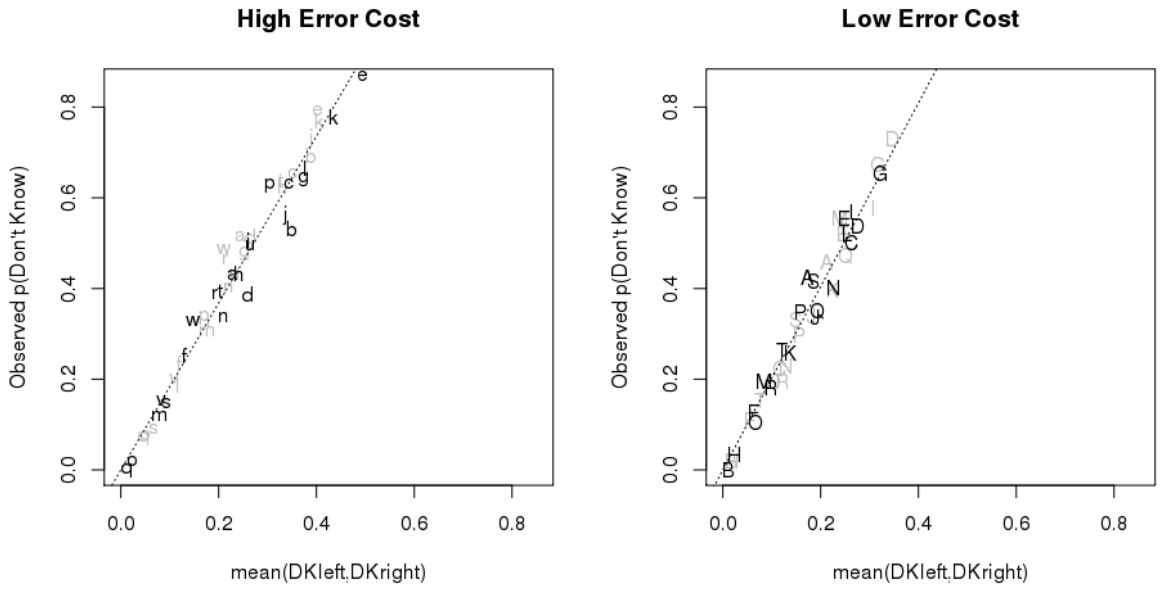
Model Parameters

In this section we focus on the model parameters of most interest, mean rates (*v*), choice thresholds (*b*) and *DK*. Results for the remaining parameters (i.e., non-decision time, t_0, start-point variability, A, and rate variability, *sv*) are presented in supplementary materials.

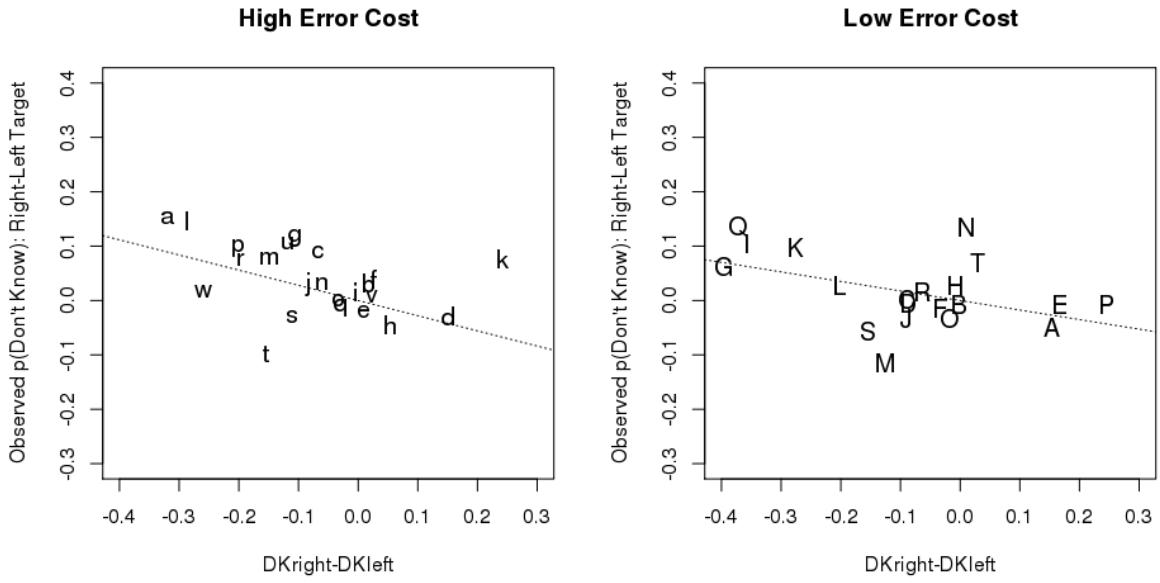
Figure 6a and 6b shows the relationship between the probability of a don’t-know response and *DK* parameter estimates. As we cannot distinguish don't-know responses corresponding to DK values for the left and right accumulators we averaged them in the plot. The figure shows a very strong linear relationship between the average DK and don't-know probability, which does not differ for speed and accuracy conditions, but which is slightly

steeper for low error cost (slope = 2.02, r^2 = .992 p < .001) than high error cost (slope = 1.84, r^2 = .993, p < .001). The corresponding regression lines, assuming a zero intercept, are superimposed on the scatter plots in Figure 6a and 6b. Although the relationship cannot continue to be linear for larger values of DK (as DK is bounded above by one) these results do show that there is a very tight relationship between DK estimates and don't-know use.

We reasoned that the difference in don't-know use for left and right targets might, at least in part, be mediated by the difference in DK for left and right accumulators. As the left accumulator would most often win for left targets DK for the right accumulator would mostly determine the frequency of don't-know responses, and vice versa for right targets, the relationship would be expected to be negative. As shown in Figures 6c and 6d this expectation was borne out, with a stronger relationship for high cost (slope = -0.28, r^2=.33, p = .004) than low cost (slope = -0.175, r^2=0.25, p = .02), where again we assumed a zero intercept.



(a) (b)



(c) (d)

Figure 6. (a) and (b): Observed probability of don't-know responses for speed (grey letters) and accuracy (black letters) as a function of the average of DK estimates over left and right accumulators. (c) and (d): Observed probability of the different between DK responses for left and right targets as a function of the difference in DK estimates between right and left accumulators. Letters a ... w correspond to the 23 participants in the high error cost-condition and A … T to the 20 participants in the low error-cost condition.

Figure 7a and 7b show average *DK* estimates, which were clearly larger for high than low error cost (*p*s < .001), consistent with more frequent don't-know responses in the former condition. *DK* estimates were also larger and for the left than right accumulator (*p*s < .001), except in the low error cost accuracy condition ($p = .23$), consistent with more frequent don't-know responses for right than left targets. In the high error-cost condition there was no support for a difference between speed and accuracy in *DK* for either the left ($p = .15$) or right ($p = .45$) accumulators. For the low error-cost, *DK* for the left accumulator was larger for speed than accuracy ($p < .001$) and vice versa for the right accumulator ($p < .001$).

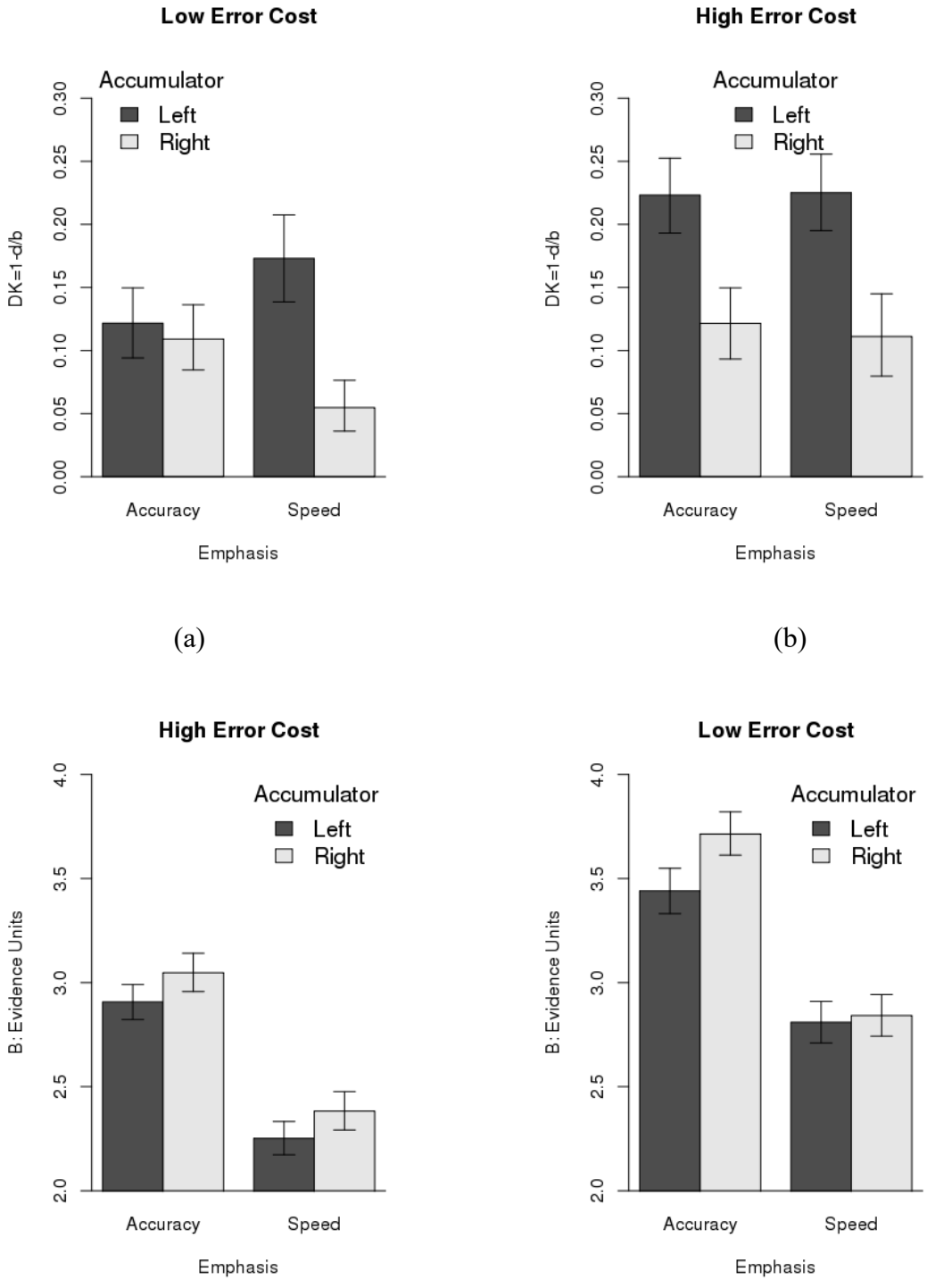


Figure 7. Threshold estimates with 95% credible intervals for left and right accumulators in accuracy conditions (a) *DK* for high error cost (b) *DK* for low error cost, (c) *B* for high error cost and (d) *B* for low error cost.

Figure 7c and 7d show that for the choice threshold the largest effect was a much higher value under accuracy than speed instructions, as expected (*p*s < .001). There was a weaker but reliable tendency for a bias to left responses (i.e., a lower threshold for the left accumulator) (*p*s < .001) except in the low error-cost speed condition (*p* = .21). Choice thresholds were also much higher under low than high error cost (*p*s < .001), consistent with faster RT in the latter condition.

Figure 8 shows that mean rates were always higher for matching than mismatching accumulators (*p*s < .001), and that the difference between them was clearly greater in the accuracy than speed instructions (*p*s < .001) except in the low error-cost condition for lower similarity pairs where the same trend was present but weaker (*p* = .29). This was mainly due to the match rate being clearly less in speed than accuracy (*p*s < .01), with mismatch accuracy greater than speed in low error cost for both high and low similarity (*p* < .01), whereas there was a weak trend in the opposite direction for high error costs for both high (*p* = .1) and low (*p* = .4) similarity pairs. Figure 8 also shows no evidence for an effect of pair similarity on rates, consistent with the lack of effect of this factor on observed response probability and RT.

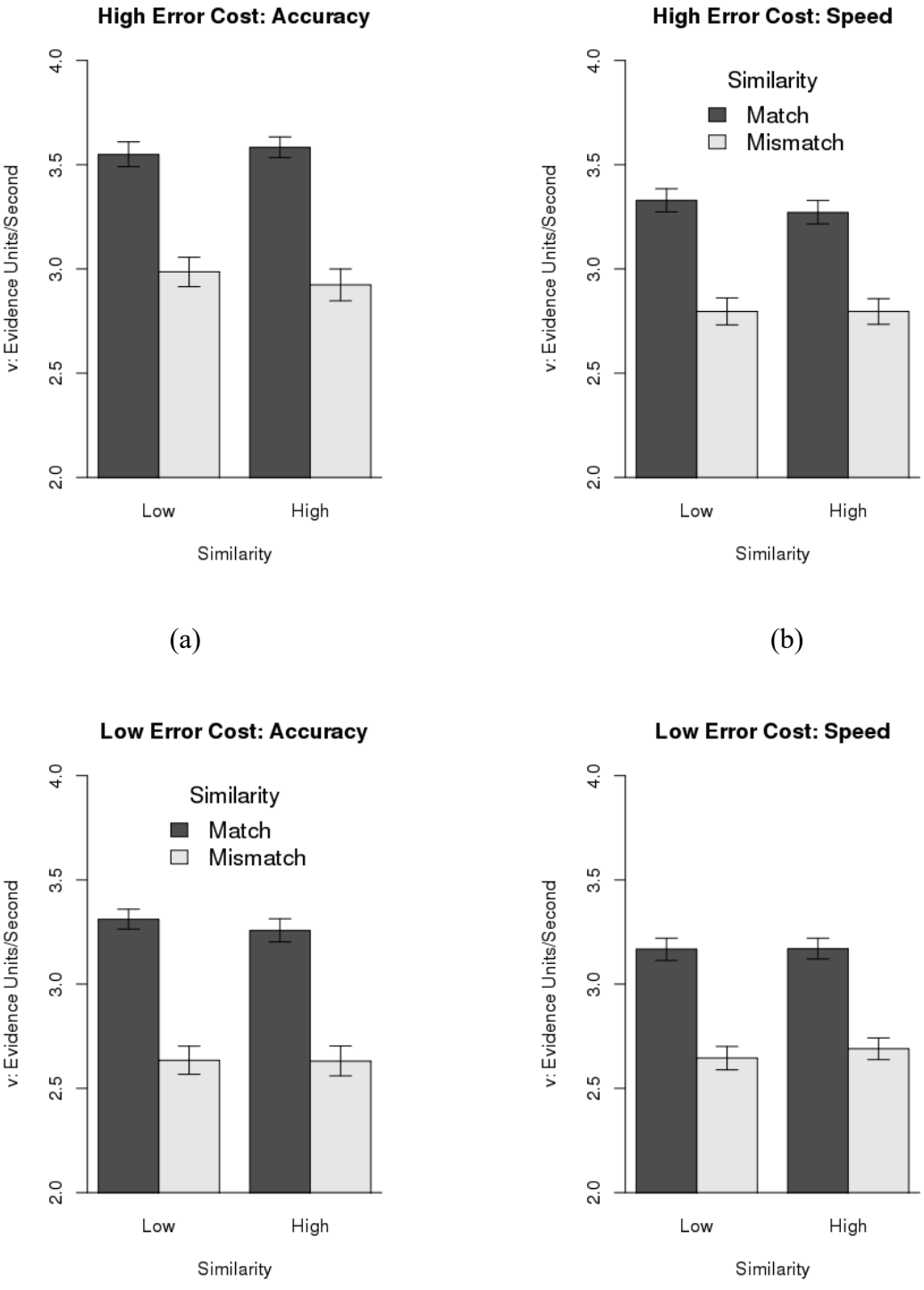


Figure 8. Mean rate (*v*) estimates with 95% credible intervals for matching and mismatching accumulators for low and high similarity pairs in (a) … (d) low/high error cost x speed vs. accuracy emphasis.

Discussion

The MTR model provided an accurate account of most aspects of Experiment 1, including the marked level of individual differences in don't-know use. These differences correspond directly to the model's *DK* parameter, which quantifies the proportion of the region under the choice threshold that produces a don't-know response when that accumulator is the loser. Although not as strong, there was also a relationship between the difference in *DK* between accumulators and the differences between the frequency of don't-know responses when targets were on the left vs. when targets were on the right. The MTR model was not only able to account for most effects on don't-know frequency but also for effects on don't-know RT, such as overall slower don't-know responses under high than low error cost and the interaction between speed-accuracy and high-low cost conditions on don't-know response speed. These manipulation effects were explained by a simple pattern of effects on DK parameters; in all cases larger DK for high than low error cost, and in most cases larger DK for the left than right accumulator. There was no evidence for a difference in DK as a function of speed vs. accuracy emphasis when error cost was high, whereas when error cost was low it interacted with the left-right difference, which largely disappeared under accuracy emphasis.

The only somewhat unsatisfactory aspect of the model's fit to don't-know response probability was an inability to account for an interaction between pair similarity and the increase from low to high cost, which was greater for lower similarity pairs than for higher similarity pairs (~ 10% vs. 6%) whereas the model estimated the same difference (~8%, see Figure 4b vs 4c). It is conceivable that this difference occurred because the greater similarity of higher than lower similarity pairs may be immediately obvious (e.g., compare Figures 2b and 2c), allowing participants sufficient time to set different don't-know thresholds for these

two types of pairs. In a related example, Provost and Heathcote (2015) found that in a mental rotation task requiring a matching decision about whether one image is a rotated version of another image presented next to it, participants set different LBA thresholds depending on the degree of rotation, which could be easily assessed regardless of whether or not the images matched. In the present data, models where don't-know thresholds were allowed to differ with similarity did improve fit but were not favoured by the DIC model selection criterion (Spiegelhalter, Best, Carlin & van der Linde, 2002) in both low (9960 vs. 9890 for the model with differing thresholds and high (10728 vs. 10678) error-cost conditions (lower values of DIC indicate a better trade-off between goodness-of-fit and model complexity).

Apart from the small interaction just described, pair similarity did not appear to have any effect either in manifest performance measures or in MTR parameters. Although it was anticipated that discounting of features shared between pair members would have some protective effect against the deleterious effects of increased pair similarity the finding that it completely removed the effect was unexpected. Unfortunately, this lack of effect, and the possibility that participants may be able to adjust don't-know thresholds as a function of similarity, meant that Experiment 1 did not afford as strong a test of the MTR model as hoped. Experiment 2 remedies both of these shortcomings. Because test items are presented one at a time it is not plausible that participants could set different don't-know thresholds for high and low similarity lures. As we now show, testing single items also resulted in lure similarity having a very large effect on performance, so together these two attributes afforded an even stronger test of the MTR model than Experiment 1.

Experiment 2

Method

Participants

56 participants were run in order to fulfil (after exclusions described in the results section) our aim of 24 participants in each between-subject condition. All attended a single one-hour session, working singly or in pairs in separate carrels, and were University of Tasmania students completing either first- or second-year psychology courses; as well as a parent of one of the student volunteers who wished to participate. Ages in years ranged from 18 to 66. The final sample had 20 participants who were administered the speed condition first (36 assigned to accuracy first), and 26 participants who were administered with left keys= target (30 assigned with right key = target). Consent and ethical procedures were as for Experiment 1.

Procedure

The procedure was the same as for Experiment 1 with the following exceptions. The test list consisted of 25 lone faces presented in a random order. The test items consisted of 12 targets, 6 lower-similarity lures and 6 higher-similarity lures randomly selected to match the middle 24 study items and one target drawn from the remainder of the study list. Participants pressed the “z” key to indicate that a target and the “/” key to indicate a lure, or vice versa, with the key mapping alternated for each new participant in the experiment.

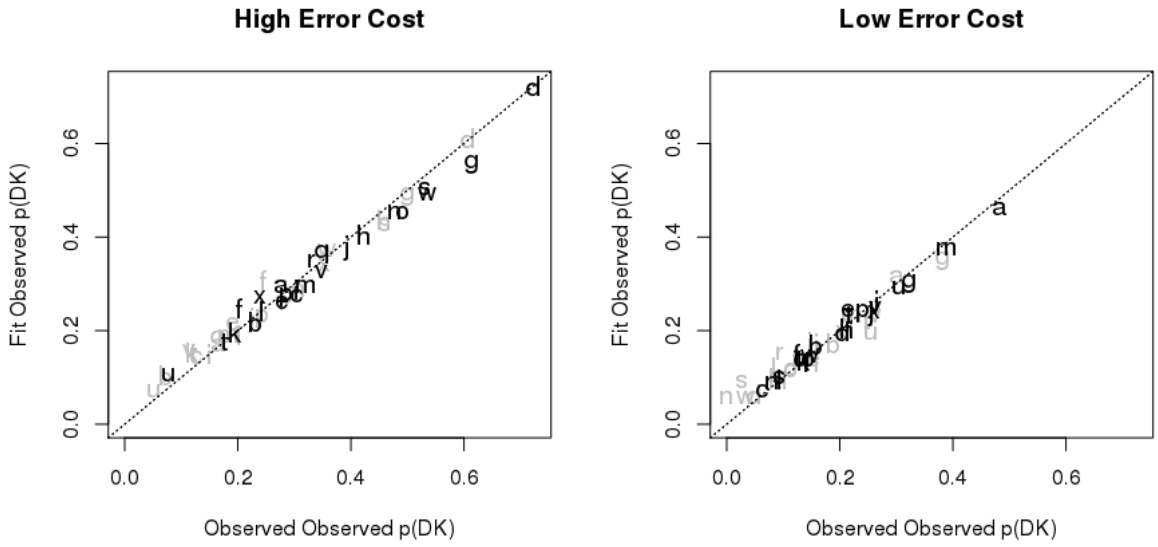
Results

Four participants were excluded because they had fewer than 4% DK responses (0.7%3.8%), three participants were excluded because, within either speed or accuracy

blocks, or both, the error rate was greater than 60%. One participant who failed to respond on 36% of trials was also excluded. The remaining 48 participants failed to respond on 0.4% of trials on average and had fast outlying responses less than 0.4s, which were removed from further analysis, on 0.96% of trials.

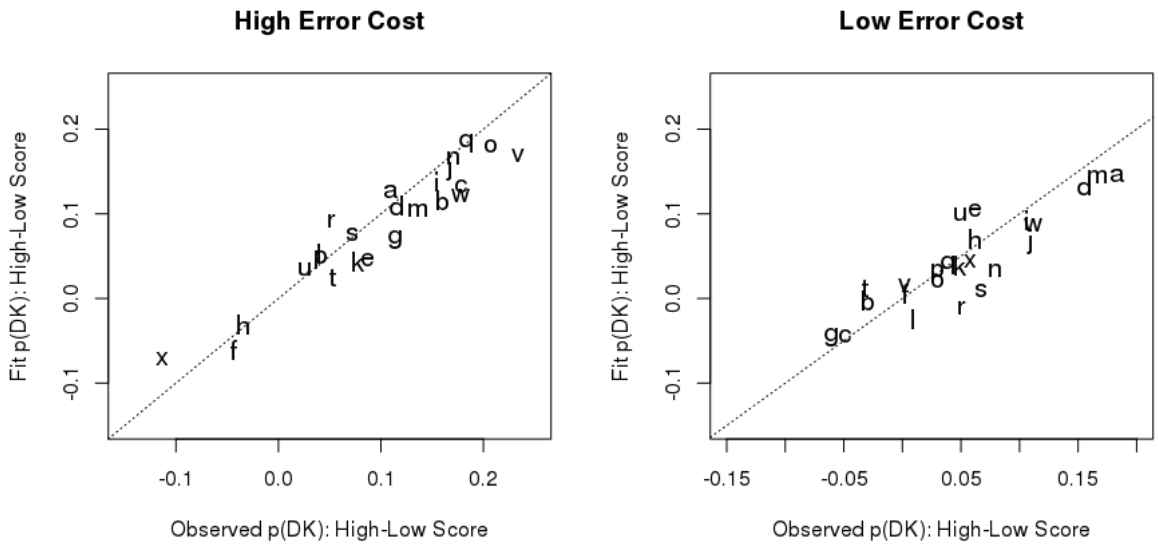
Analysis methods were the same as for Experiment 1. Because lure similarity had a strong effect, for brevity we refer to low and high similarity lures as easy and hard stimuli respectively. In the course of our preliminary analysis we realized that participants were using don't-know responses differently depending on their current game score. Recall that in order to maintain motivation we did not allow scores to go below zero, meaning it is no longer necessary to use the don't-know response to protect against losing points. In Experiment 1 this virtually never happened because overall performance was fairly accurate but Experiment 2 was much more difficult. Consequently, their score frequently fell to zero or close to zero, and it was clear that don't-know use was then reduced, particularly in the high error-cost condition. Further, at a score of 100 points or less there is no difference between the high and low error-cost conditions, in that the same amount was lost for an error in both. We therefore created a factor that divided trials into low (100 points or less) and high scores. With this cut-off 38% of trials had low scores overall (51% for high error cost and 25% for low error cost). We included the score factor in our analysis of response probability and men RT and into the MTR parameter specification for the don't-know threshold, instantiating the assumption that score had a simple selectively influence.

As for Experiment 1, the MTR model had 19 parameters in total[3]. As in Experiment 1, there were four choice threshold (b) parameters (estimated as B), one for the old and one for the new response accumulator separately for the speed condition and for the accuracy condition. Don't-know thresholds were estimated as a proportion of the choice threshold (i.e., we estimated 1-DK directly), with these proportions allowed to vary over score and response. This meant that there were 4 estimated relative-threshold parameters which give 8 different *d* threshold values, since the model estimates separate threshold heights for *b* under speed/accuracy emphasis. There was one start-point noise (*A*) and one non-decision time (t_0) parameter. The *sv* parameter for the mismatching accumulator was fixed at 1 for each stimulus type (easy or hard lures and targets) and estimated for the matching accumulator for each stimulus type. There were 6 mean rate parameters, half for the matching and half for the mismatching accumulator, with different values for high and low similarity lures and targets.



[3] We also explored models with more parameters and found one with 29 parameters that was preferred by DIC, but which did not provide a noticeably better description of the data than the simpler model. We report details of the more complex model in supplementary materials.

(a) (b)



(c) (d)

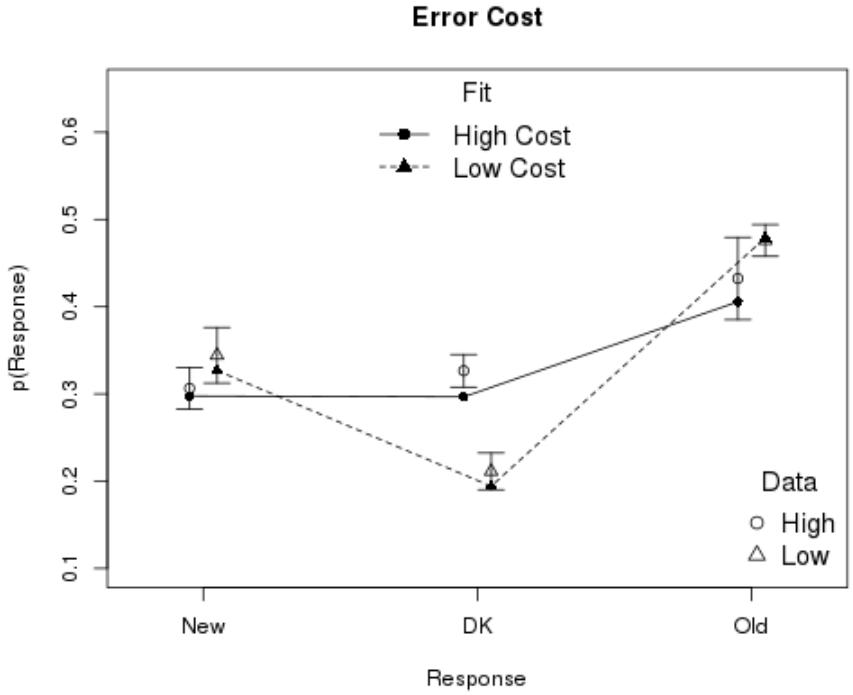
Figure 9. (a) and (b): Observed vs. fitted don't-know probability for low-score trials (grey letters) and high-score trials (black letters). Letters correspond to participants: a ... x for participants in the high error-cost condition and A … X for participants in the low error-cost condition. (c) and (d) observed and fitted differences between high and low score trials results.

For details of sampling methods and priors see supplementary materials, with data files and R code to fit the models are available at osf.io/6h4qe/. In supplementary materials we plot the global fit of the model to all conditions, both in terms of RT distribution (as represented by the 10th, 50th and 90th percentiles and cumulative distribution functions) and response probability. As for Experiment 1, these plots demonstrate the model provides an accurate description of the data.

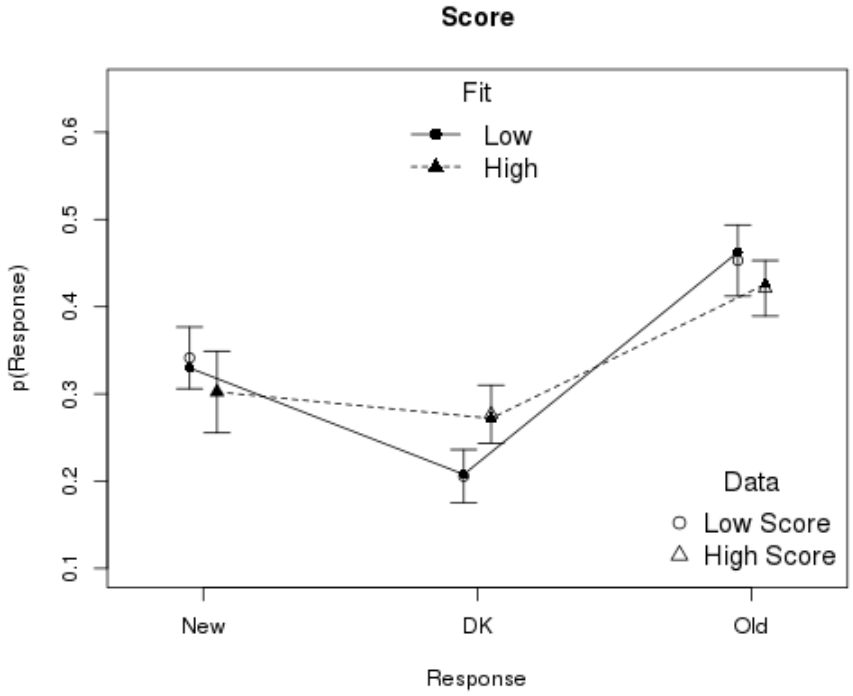
Figures 9a and 9b show that we again observed very large individual differences in don't-know use, which was on average greater for high than low error cost, and that the MTR model was again able to accommodate these findings. Figures 9c and 9d show that for almost all participants in the high error-cost condition, and for many in the low error-cost condition,

don't-know use was much more frequent when the score was high, although in one case (participant “x”) this was reversed.

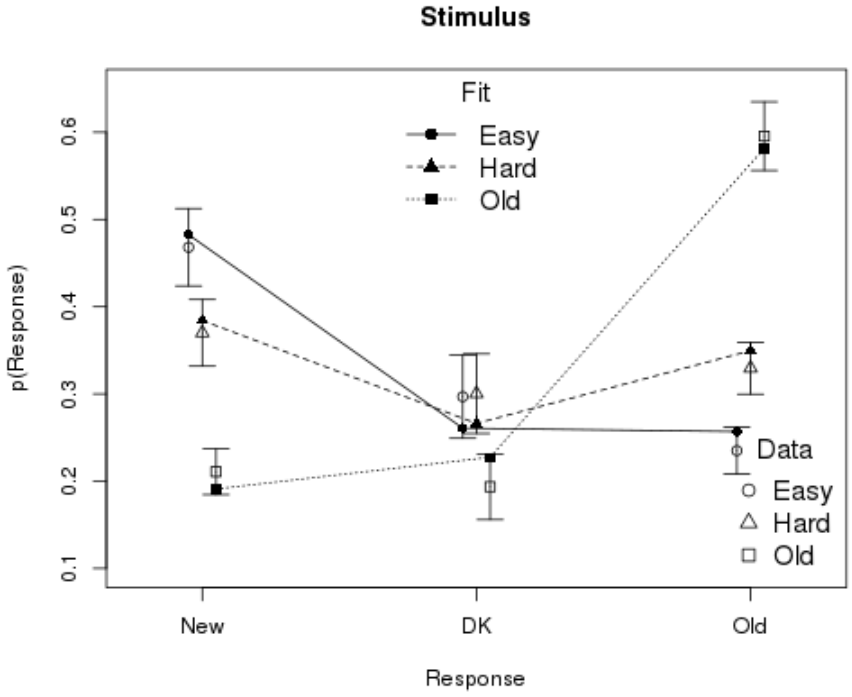
Figure 10 shows that on average the probability of a don't-know response was reduced when the error cost, $\chi^2(1) = 13.4$, $p < .001$, or score, $\chi^2(1) = 95$, $p < .001$, were low, and also for target test items compared to easy and hard lures, $\chi^2(2) = 216$, $p < .001$, which had almost identical don't-know probabilities. No other effects on don't-know probability were significant. The model captured these results fairly well, except for under-predicting high error cost in Figure 10a and the magnitude of the lure vs. target difference in Figure 10c. No other effects on don't-know probability were significant, including speed vs. accuracy emphasis, as shown in Figure 10d. Figure 10c shows a substantial effect of lure similarity on accuracy; accuracy was barely above chance for hard lures (52.9%), intermediate for easy lures (65.5%) and best for targets (74.2%), $\chi^2(2) = 326$, $p < .001$. As shown in Figure 10d, there were also more correct responses under accuracy emphasis (69.4%) than speed emphasis (66.5%), $\chi^2(1) = 9.2$, $p = .002$. Both effects were well captured by the model. No other effects on the accuracy of definitive responses were significant.


Error Cost
Fit
High Cost
Low Cost
p(Response)
0.1
0.2
0.3
0.4
0.5
0.6
Data
High
Low
New
DK
Old
Response

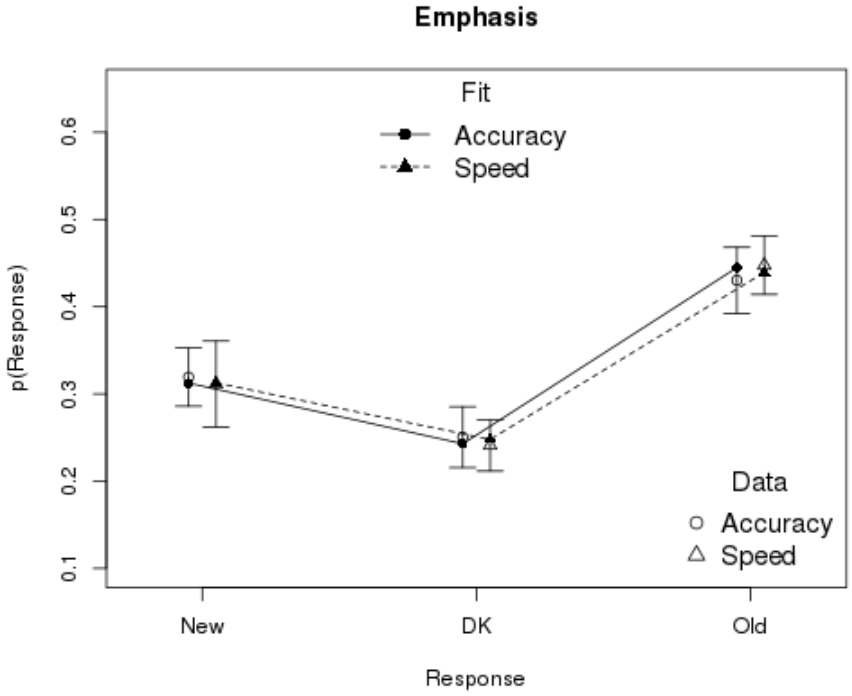
(a)


Score
Fit
Low
High
p(Response)
0.1
0.2
0.3
0.4
0.5
0.6
Data
Low Score
High Score
New
DK
Old
Response

(b)

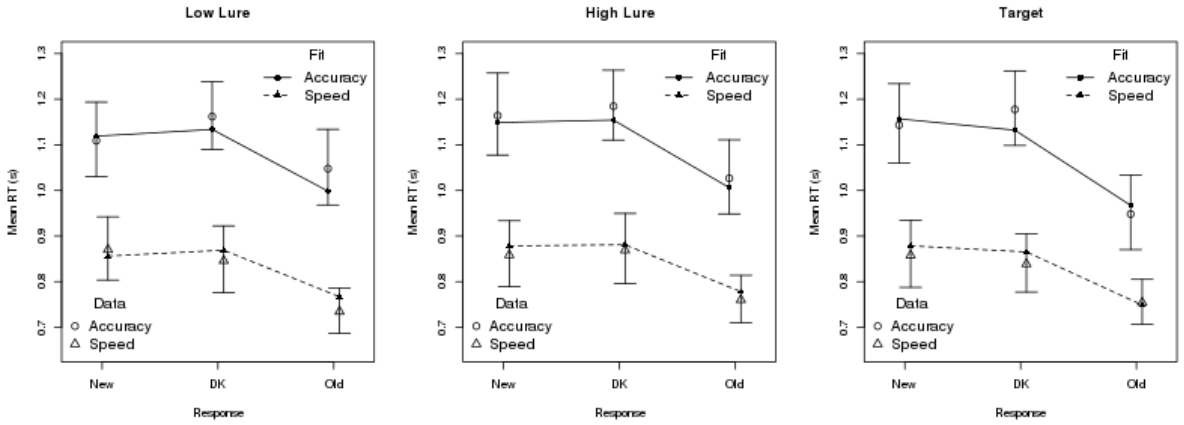


(c)

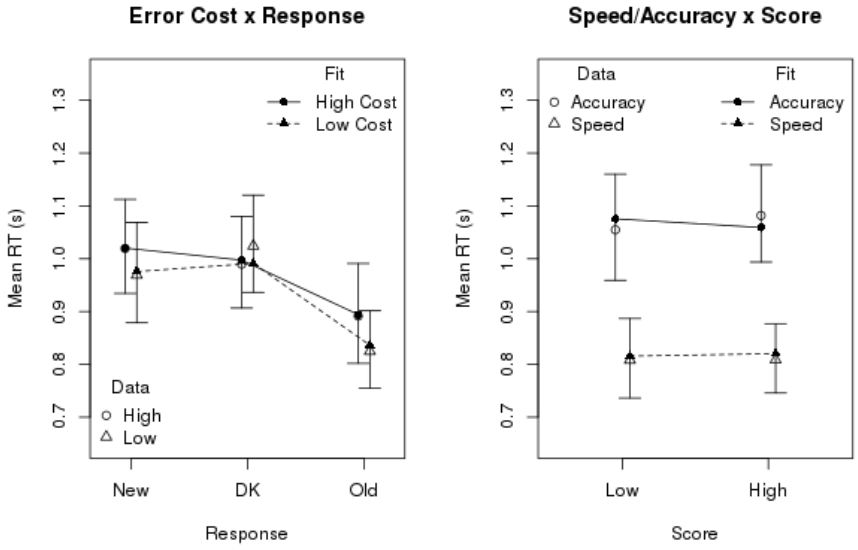


(d)

Figure 10. Response probability for data (open symbols with 95% confidence intervals) and fits (closed symbols joined by lines) as a function of response type and (a) error cost, (b) score, (c) stimulus and (d) speed vs. accuracy emphasis.



(a)



(b) (c)

Figure 11. Mean RT for data (open symbols with 95% confidence intervals) and fits (closed symbols joined by lines). Results as a function of response: (a) broken by speed vs. accuracy emphasis for each stimulus and (b) broken down by error cost. (c) Results as a function of score broken down by speed vs. accuracy emphasis.

Mean RT was faster in speed than accuracy emphasis by ~0.26s, $\chi^2(1) = 4325$, $p <$ .001, target responses were faster than lure and don't-know responses by ~0.13s, $\chi^2(2) = 942$, $p <$ .001, responses to targets were faster than responses to lures by ~0.06s, $\chi^2(1) = 27.2$, $p <$ .001, and low-score trial responses were faster than high-score trial response by ~0.03s, $\chi^2(1) = 16$, $p <$ .001. Figure 11a shows that there was a significant three-way interaction in mean

RT between response, stimulus and speed vs. accuracy, $\chi^2(4) = 31.9$, $p < .001$ (all constituent two-way interactions were also significant). For don't-know responses the emphasis effect was larger for targets than lures by ~ 0.05s. For lure responses it was smaller for easy lures than hard lures and targets by ~0.6s. For target responses it was smallest for targets by ~0.08s from hard lures and was bigger again by ~0.04s for easy lures. As shown in Figure 11b, there was an interaction between error cost and response, $\chi^2(1) = 40.9$, $p < .001$, with slowing for high over low cost for definitive response by ~0.05s on average, but speeding for don't-know responses, again by ~0.05s. Finally, as shown in Figure 11c, there was a **small** two-way interaction due to a larger emphasis effect for low than high score trials by ~0.03s, $\chi^2(1) = 15.6$, $p < .001$. No other effects were significant. The model did a good job of accommodating all of these effects except the later small two-way interaction.

Model Parameters

Once again, we focus on the mean rates (v), choice thresholds (b) and DK in this section, with results for the remaining parameters presented in supplementary materials.

Figures 12a and 12b show the association between DK and the probability of a don't-know response. As for Experiment 1 there was a strong linear relationship with a slope close to two for both high error cost (slope = 1.9, $r^2 = .99$, $p < .001$) and low error cost (slope = 1.83, $r^2 = .97$, $p < .001$). As shown in Figures 12c and 12d, the same type of strong linear relationship held between the difference in don't-know probability between high and low trials for both high error cost (slope = 2.2, $r^2 = .96$, $p < .001$) and low error cost (slope = 2.1, $r^2 = .87$, $p < .001$) conditions.

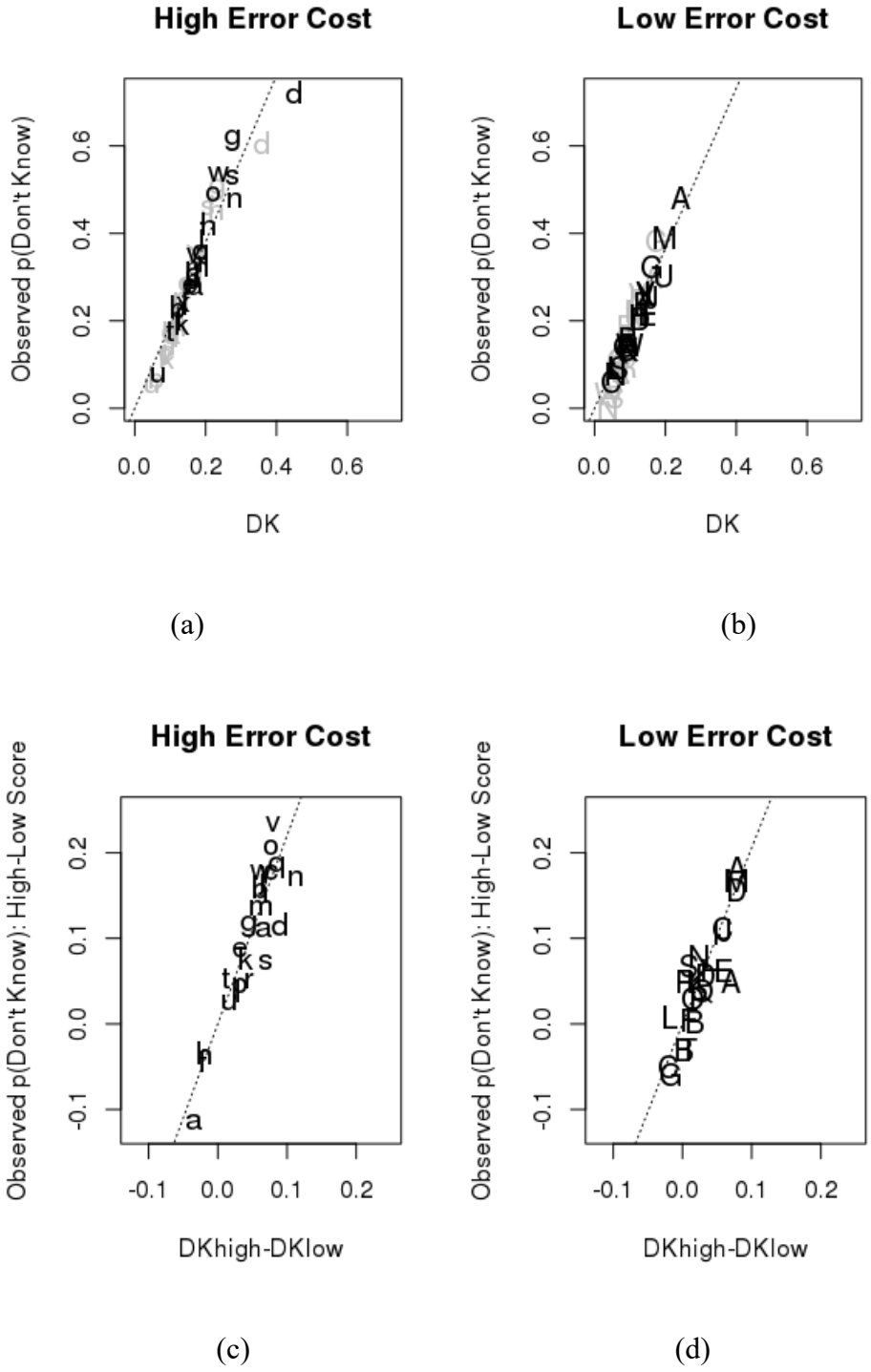


Figure 12. Observed probability of a don't-know responses for low-score (grey letters) and high-score (black letters) trials as a function of the average of DK estimates over lure and target accumulators. (c) and (d): Observed probability of the different between don't-know responses for high and low score trials as a function of the difference in DK estimates between right and left accumulators. Letters a ... x correspond to participants in the high error-cost condition and A … X to participants in the low error-cost condition.

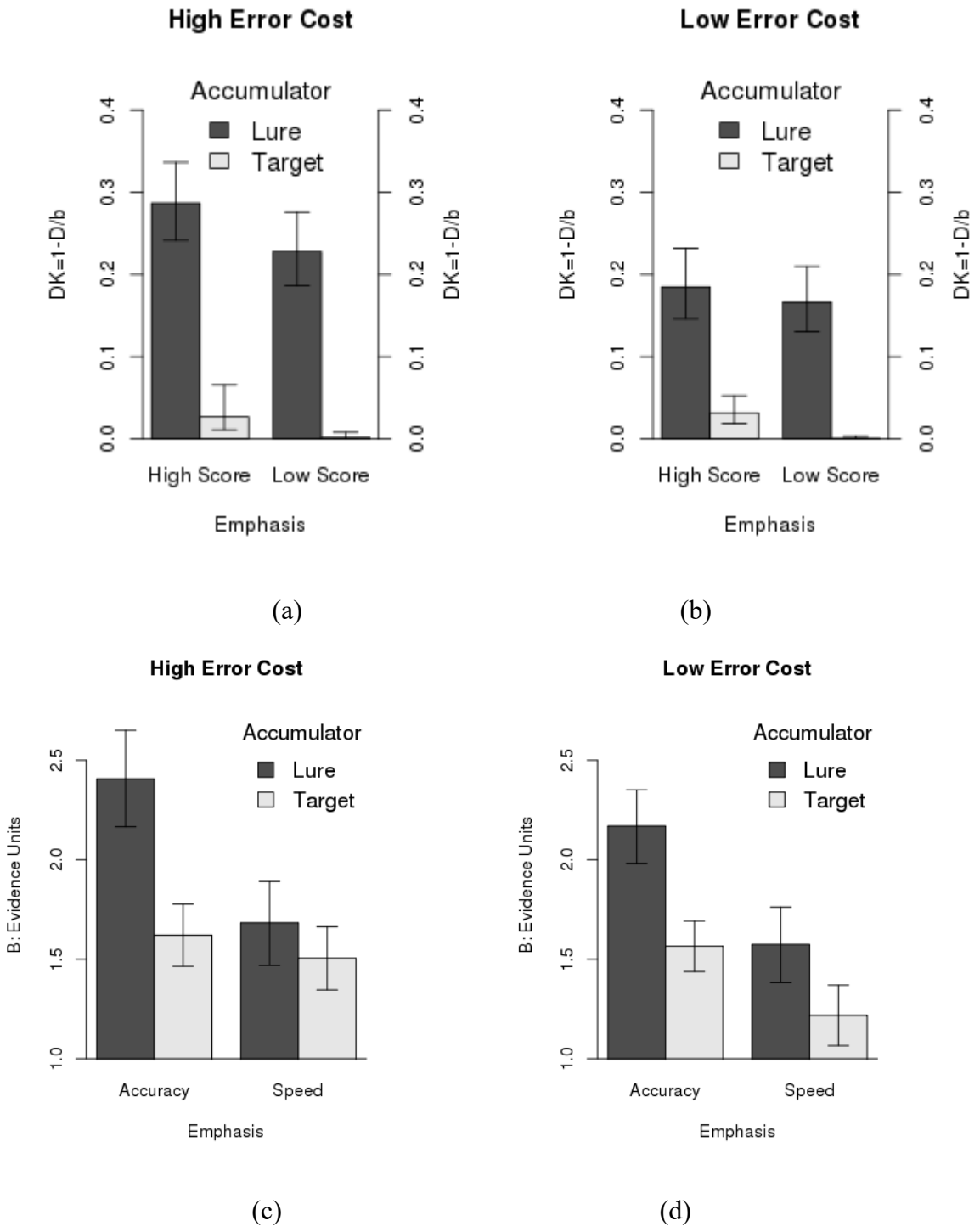


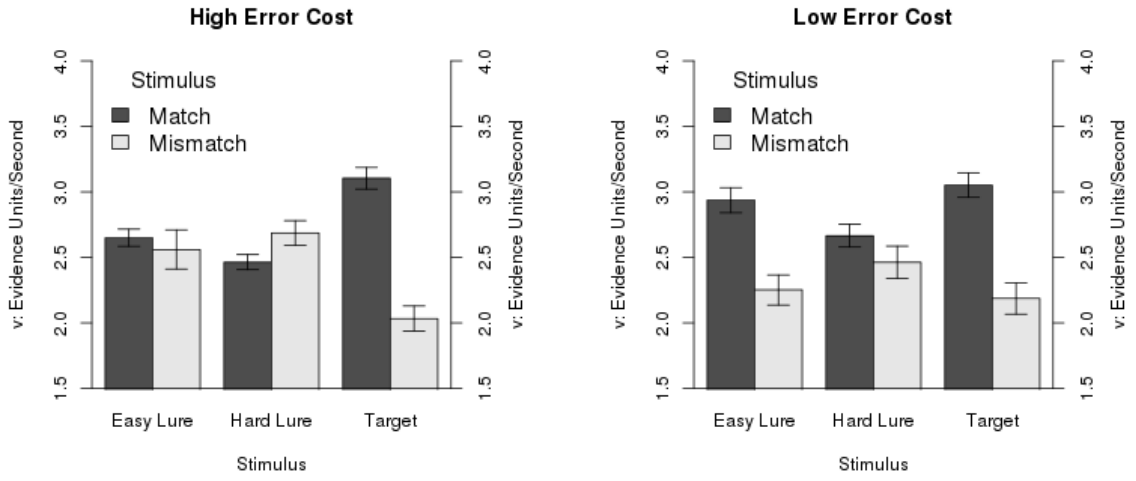
Figure 13. Threshold estimates with 95% credible intervals for lure and target accumulators. (a) – (b) *DK* estimates; (c) and (d) choice threshold *B* estimates.

Figure 13 shows that *DK* estimates in the high error-cost condition were larger for high than low score trials (*p*s < .001). DK estimates are smaller when on a Low Score than a high score, with a bigger difference in the High Error cost condition. Lure DK estimates were

much bigger than target DK estimates, suggesting that responses that resulted in the lure accumulator winning the choice race rarely became don't know responses.

Choice threshold estimates were greater under accuracy than speed emphasis (*p*s < .001) and were greater for the lure than target accumulator (*p*s < .001). The difference between target and lure is bigger in the Accuracy emphasis condition.

Figure 14 shows that in most cases the match-accumulator rate was greater than the mismatch-accumulator rate (*p*s < .001), except for hard lures in the high error cost condition. Target stimulus had the biggest matching drift rate, smallest mismatching drift rate and subsequently the biggest difference between match and mismatching accumulators (ps<.001). Easy lure stimulus trials had bigger matching and smaller mismatching mean drift rates than hard lure stimulus (*p*s<.01). The difference between match and mismatching accumulators is biggest on target accumulators in the high error cost condition, and bigger for lure stimulus in the low error cost condition.



(a) (b)

Figure 14. Mean rate (*v*) estimates with 95% credible intervals for matching and mismatching accumulators for easy and hard lures and target stimulus trials.

Discussion

As for the two-alternative forced choice performance in Experiment 1, the MTR model provided a quite accurate account of most aspects of single-item recognition in Experiment 2. Once again there were marked individual differences in don't-know use that had a direct relationship with the MTR's average *DK* parameter, which corresponded to about half of a participant's probability of a don't-know response. There were also more subtle and varied effects of the within- and between-participant manipulations on *DK* estimates, which were generally larger when the cost of an error was greater and larger for the lure than target accumulator.

Just as in Experiment 1, larger *DK* values when there was a high error cost translated into more frequent don't-know use. Greater difficulty in Experiment 2 meant that participants had very low scores on half of the high error-cost trials and one quarter of the low error-cost trials, and for most participants this was associated with decreased don't-know use. The MTR model was able to accommodate this effect by a simple reduction in *DK* for the lower-score trials. There was also a clear 1:2 relationship between *DK* estimates and don't-know response probability for lower minus higher error-score trials that explained individual differences in the effect of score, including reversals of the usual pattern for some participants. This strong and direct relationship contrasts with the less direct relationship between *DK* estimates and the probability of don't-know responses for left vs. right targets in Experiment 1, which would be expected to be weaker because there is only a partial correspondence between target side and the accumulator that determines whether a definitive or don't-know response is made.

These findings, and those for individual differences in the effect of target side in Experiment 1, illustrate that although it is sometimes possible for some manipulations and

individual differences to obtain approximate estimates of *DK* from don't-know response probability, a full fit of the MTR model can be required in order to evaluate how *DK* values change for other effects.

In contrast to Experiment 1, the similarity manipulation in Experiment 2 was successful, with lower (indeed, close to chance) accuracy for lures that were highly similar to studied items relative to lures that were less similar. Although don't-know use did not differ between easy and hard lures, it was higher for lures than for targets, which produced faster and more accurate responses than lures. The MTR model was able to accommodate this pattern of don't-know use, but it slightly under-predicted the lure-target difference. Greater don't-know use on target stimulus trials was predicted even though the model predicts that DK was much bigger on the lure accumulator than the target accumulator, implying that most don't know responses would have been target responses. This means that most don't know responses on target stimulus trials would have been correctly identified, and that most don't know responses on lure stimulus trials would have been errors. With the don't know responses being predicted mainly from one response, the model is not able to fully explain the difference in don't know use by stimulus.

Once again this illustrates that don't-know use is governed not only by the strategic factors that determine the setting of evidence thresholds, but also by other factors, such as those associated with the stimuli on which choices are based.

As expected, accuracy in the speed emphasis condition decreased because both choice thresholds were lower. In the supplementary materials we include a more complex fit that also allows the drift rates to vary by the speed emphasis condition, which predicts a reduction in evidence quality in addition to the lower thresholds.

We also found two types of bias: *response bias*, which is mediated by choice thresholds, and *stimulus bias*, which affects evidence accumulation rates (e.g., Leite &

Ratcliff, 2011; Osth, Dennis & Heathcote, 2017; White & Poldrack, 2014). First, for choice thresholds there was a consistent response bias towards target responses (i.e., lower choice thresholds for the target than lure accumulator) that was larger under accuracy than speed emphasis. There was also a stimulus bias towards old responses which was so extreme in the high error cost condition, that a hard stimulus provided a bigger drift rate for target responses than lure responses on average.

In summary, since the hard lures were highly similar to target responses, there was both a stimulus bias and a threshold bias towards identifying the stimulus as a target. However, since DK was much larger on lure than the target accumulator, if a target response was not a clear winner, it would become a don't know response instead. That participants are so adept in taking advantage of the flexibility afforded by the availability of a don't-know response option suggests that the strategy of deciding not to choose is one that they have experience with because of its utility in dealing with difficult choices.

General Discussion

In this paper we proposed a theory of what have been variously called "equivocal", "uncertain", "opt-out"—and in our terminology "don't-know"—responses as a third option when faced with a decision as to which of two choices is correct. We instantiated the theory in a dynamic evidence-accumulation model, the Multiple Threshold Race (MTR), that can explain how often each of the three possible responses is made and the distribution of corresponding RTs. Don't-know responses have been shown to be used adaptively when humans and other higher mammals (Shields, Smith & Washburn, 1997; Smith et al., 1995) are faced with difficult choices and to improve performance with line-ups in eyewitness memory (Weber & Prefect, 2012). We also found that our participants were adept at the use of don't-know responses.

We applied the MTR to recognition memory paradigm with face stimuli where the decisions were difficult because of a high degree of similarity between faces that had (targets) and had not (lures) been previously studied. We found adaptive behaviour in the sense that participants strategically avoided a loss of points associated with an error by making a no-penalty don't-know response. They did so when the level of error cost was constant across the entire experiment and, displaying flexibility that we had not anticipated, they also modulated don't-know use on a trial-by-trial basis when their error cost decreased because their points tally was low and could not decrease below zero. We also found that the availability of don't-know responses enabled participants to adapt to instructions emphasising either the speed or accuracy of responses in order to deal with one choice stimulus (lures) being particularly difficult. They did this via coordinated changes in not only the threshold controlling the level of don't-know responding but also by modulating two other aspects of the decision process: response bias and stimulus bias (White & Poldrack, 2014).

These findings add to a growing catalogue of examples (e.g., Palada et al., 2016, 2018, in press; Ratcliff & Rouder, 1998) where participants strategically modulate a variety of aspects of the evidence-accumulation process in order to cope with difficult decisions. As in these other examples, we found that a model based on the idea of evidence-accumulation to a choice threshold provided both a good descriptive account of the probability and speed of definitive choices when faced with such challenging choices. By simple addition of an extra don't-know threshold to the evidence accumulators corresponding to each binary choice we were able to extend this framework to provide a good description of the probability and speed of don't-know responses. As was also the case in the other examples, the parameters of the extended model provide insights into the psychological mechanisms by which participants managed the demands of a difficult choice scenario that would not be possible by examining only their observed behaviour. These insights reinforced the case for increased study of how

decisions use don't-know responses and potentially for greater use of don't-know responses in decision making applications given the apparent flexibility and adaptability with which they were used by our participants (see also Weber & Perfect, 2013).

In the past, allowing don't-know responses has not been favoured for two reasons. The first is a high level of individual differences in their uptake, which has the potential to add noise that obscures measurement of the underlying ability to make an accurate choice. We also found very marked individual differences. Some participants made little use of don't-know responses despite being shown that their decisions were highly error prone and being encouraged to use them as a means to avoid penalties associated with errors. Other participants used them so often that they missed opportunities where they would have likely been rewarded for making a correct response. However, the MTR was able to accurately accommodate these differences in terms of the position of the additional don't-know threshold. Unexpectedly we found a very simple relationship, whereby an individual's probability of making a don't-know response corresponded to half of the average proportion of the region under the choice threshold above the don't-know threshold. This relationship also extended to modulations of the don't-know threshold for low vs. high scores, but for other effects, such as the difference between targets presented on the left vs. right of a two-alternative forced choice and the effect of speed vs. accuracy emphasis, the relationship was more complex so that fitting of the MTR model was required.

The second reason against allowing don't-know responses is the conviction that forcing participants to make a binary choice is more informative about a participant's ability. Watson et al. (1973) presented empirical evidence against this contention in the context of static signal detection theory of response probabilities. They showed that the same measures of discrimination ability were obtained as when making a binary choice, as well as accommodating individual differences, by simply estimating an additional response criterion.

We took a similar approach but added two thresholds to provide an account of both RT and choice probability. In supplemental materials we demonstrate that the MTR models we reported here had good measurement properties in the sense that they provided accurate and precise parameter estimates when fit to simulated data for the same designs as used in our experiments. These results suggest that, by using the full information available in response probabilities and RT, the MTR model is able to provide valid and informative estimates of not only the extra thresholds but all of the other model parameters that provide a rich and informative characterization of the psychological process of decision making.

However, we also show in supplemental materials that this may not always be the case. In particular, we found in the design used in Experiment 1 that if we allowed extra flexibility by estimating different rates for stimuli with different target locations, parameter recovery failed due to extreme tradeoffs between rate and don't-know threshold parameters whereby don't-know thresholds for one accumulator were reduced to zero. This appears to be a result of uncertainty intrinsic to the MTR about which accumulator's choice threshold triggers a don't-know response. Hence, we advise that applications of the MTR (and indeed any cognitive model) be accompanied by a parameter-recovery study (Heathcote, Brown & Wagenmakers, 2015; Heathcote et al., 2018). The need for this check is reinforced by an examination, detailed in supplementary materials, if how adding a don't-know threshold affected measurement of other model parameters. This examination revealed that when response biases exist (such as in the first experiment), there can be regions of unidentifiable parameters, particularly in complex models. In particular, when the drift rates vary by stimulus type and thresholds by accumulator in the design of Experiment 1, non-identifiable regions are created, with the don't-know thresholds becoming biased towards one response. This creates cases were the true direction of the rate and threshold biases can flip in parameter recovery. However, this does not occur when drift rates do not vary with stimulus,

as was assumed in the model we reported. In the Experiment 2 design, there were no problems even in a more flexible model where drift rates varied with stimulus and thresholds with accumulator. This was likely the case because the large differences in accuracy and RT between stimulus types in Experiment 2.

Future Directions

Higham (2007) showed that the optimal level of don't-know use on the Scholastic Aptitude Test (a multiple-choice test with formula scoring) could be determined based an equal variance version of Type-1 and Type-2 signal detection theory. Because the MTR is a fully specified quantitative model it also enables investigation of optimality, as has been done with standard evidence accumulation models (e.g., Bogacz et al., 2006; Garton, Reynolds, Hinder & Heathcote, 2019, Starns & Ratcliff, 2010). Following the methodology developed by Garton et al., all parameters could be fixed at their estimated values while the don't-know threshold is varied in order to identify the optimum setting (and hence level of don't-know use) relative to a given payoff scheme or other loss function. This approach moves beyond previous work by taking into account RT when addressing the question of what an optimal level of don't-know use.

Although the version of MTR we applied here was largely successful, alternative versions are plausible and may be required in other paradigms. For example, in some circumstances it might be better to decide using a “threshold counting” rule, such as responding as soon as any two thresholds have been crossed. In that case, a don't-know response would be associated with crossing of the two lower thresholds first. Such a rule may be useful because accumulating more evidence to the top threshold slows responding but doesn’t change the outcome (a don’t-know response) once the two lower thresholds have crossed. As this rule results in faster don't-know responses than the rule we assumed here it

may be more suitable in cases such as the game-show scenario described in the introduction, where don't-know responses have utility only when they result in an opportunity to answer more questions in a fixed time.

The intuition behind why both rules capture uncertainty is the same; don't-know responses are associated with smaller differences between evidence totals. However, the alternative threshold-count rule represents a more radical departure from Vickers' (1979) Balance of Evidence hypothesis, which shares with the rule assumed here the assumption that only one threshold in each accumulator can trigger a response. It would seem desirable in future work to compare these two approaches, and other possibilities, such as a direct influence of knowledge of how long it takes to make each decision so as to terminate slow decisions (Kiani & Shadlen, 2009, Malhotra, Leslie, Ludwig & Bogacz, 2017). A promising process account of the latter type is to add to the usual binary evidence accumulation process a third process (see Hawkins & Heathcote, submitted) that measures the passage of time (which itself could be an accumulation process, see Simen, 2016). If the third process wins then a don't-know response is made, with the speed of the third timing processes determining the speed of don't-know responses. This approach can improve accuracy, but only if errors are usually slower than correct responses, so it remains to be seen if it is able to provide a better account than the MTR model under circumstances that require speeded responding, as it is just such circumstances that are usually associated with faster errors.

The idea of a time-out has been used to account for response deferral in the realm of value-based or preferential choices (Bhatia & Mullett, 2016; Jessup, Veinott, Todd & Busemeyer, 2009). Unlike the choices studied here, no preferential response is necessarily correct (e.g., choosing which movie to watch). Tversky and Shafir (1993) found that deferring such choices is made more likely when choice conflict is increased by enlarging or improving the choice set. Jessup et al. considered explanations of effects on deferral of

enlarging the choice set based on Decision Field Theory (Busemeyer & Townsend, 1993) in terms of three mechanism, an explicit deferral response, deferral based on a criterion number of changes in the choice with the most evidence, and deferral responses being triggered by a time limit. The latter time-out mechanism, in combination with in Bhatia's (2013) associative accumulation model (AAM), was found by Bhatia and Mullett to provide a comprehensive explanation of various ways in which deferral is affected by conflict among different attributes of multiple-attribute options, and of two effects related to deferral speed, slower deferral than definitive choices and slowing in definitive choices when a deferral response is allowed compared to when it is not allowed. The latter effects arise because slower potential definitive choices tie out and become deferrals, so the definitive choices that are actually made tend to be faster.

If deferral were explained in the same way as don't know responses, then the MTR also predicts slower deferred than definitive responses, since responses where the losing accumulator has more evidence will occur more often when the winning accumulator is slower than average[4]. It also predicts, all things being equal, that definitive choices are faster when don't know or deferral responses are an option than when they are not, again because deferring or responding don't know removes what would otherwise have been slower definitive responses. It differs from the time-out account in that the possibility of deferral does not change response speed aggregated over all response types because that depends only

[4] Lowering the intermediate threshold will make don't know responses more frequent and faster on average, however it would also have the effect of making the remaining definitive responses faster again. A partial exception can occur if the intermediate threshold differs between accumulators. For example, if it is lower for the left than right accumulator, then would-be right responses are more likely to become don't know responses. This will reduce the relative number of right responses compared to left responses. If, in addition, there is also a rate bias towards right responses, so that right responses are faster than left responses, it is possible that the don't know responses can be faster than the definitive left responses.

on the choice threshold, whereas speed increases in the time-out account because time-out responses are at least as fast, and sometimes faster, than the definitive responses they replace. Future research might compare accuracy-based choice with and without don't know or deferral to see if the same slowing occurs as with preferential choice. Both in these experiments, and preferential choice experiments, the contrasting time-out and MTR predictions about overall aggregated speed might be compared.

Both mechanisms would be challenged by cases in which non-definitive responses have utility only if they are faster than definitive responses, as was suggested with respect to the game-show scenario discussed previously, perhaps necessitating something like the threshold-counting mechanism, although the time-out mechanism would still have utility due to it increasing aggregate speed. Another case in which time-outs might be preferable concerns is when non-definitive responses are made in response to the absolute rather than relative level of evidence.

White, Hoffrage and Reisen (2015) discuss both absolute and relative bases for deferring decisions in the context of preferential choice, either that no choice is good enough or uncertainty regarding which is the best. Their two-stage two-threshold model suggests that deferral can arise either during an initial absolute-evaluation stage (if no options are good enough) or during a subsequent relative evaluation stage (if good enough options cannot be discriminated). An absolute basis is less common in the accuracy-based decisions that we addressed here but does arise with respect to optional choice in eyewitness-memory research. For example, Clark's (2003) WITNESS model holds that identification decisions involve assessments of both absolute similarity (the extent to which the best matching lineup matches the witness' memory of the culprit) and relative similarity (the extent to which the best matching lineup member is favoured over the other lineup members). For a lineup member to be identified, criterion levels of both absolute match and relative superiority must be met.

Thus, an entire lineup may be rejected because no member sufficiently matches the witness's memory of the perpetrator in an absolute sense, or, when two or more members pass the absolute criterion and cannot be differentiated. The MTR model naturally addresses the relative basis but may not be sufficient for the absolute basis. In contrast the time-out mechanism is sensitive to the absolute level of evidence (i.e., if all options are weak none may accrue the required amount of evidence sufficiently quickly). Future research may manipulate the levels of absolute and relative evidence in optional choice to see if one or other mechanism, or both, or White et al.'s sequential absolute-then-relative mechanism, is favoured.

Unlike time-out models (at least without augmentation) the MTR framework can be easily extended to account for another response associated with uncertainty, multiple confidence ratings, which in contrast to don't-know responses have been studied intensively and have a range of recognized benchmark phenomena (e.g., Moran, Teodorescu & Usher, 2015). Changing the response rule in the architecture examined here, with two thresholds per accumulator, can accommodate high vs. low confidence ratings for each choice, as the likelihood for don't-know responses is simply the sum of the likelihoods corresponding to low confidence choices. More confidence ratings can be accommodated in the same way by adding extra thresholds. A similar extension is also straightforward with the threshold-counting version of the MTR. These extensions are most tractable when the choice and confidence rating are made simultaneously (Ratcliff & Starns, 2009, 2013) but given that multiple-threshold models have been used to model sequential decisions (Van Zandt & Maldonado-Molina, 2004) there is also potential to develop MTR models that address choice-followed-by-confidence-ratings paradigms (Pleskac & Busemeyer, 2010).

References

Angell, F. (1907). On the judgement of "like" in discrimination experiments. *American Journal of Psychology*, 18, 253-260. http://dx.doi.org/10.2307/1412416

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bhatia, S. (2013). Associations and the accumulation of preference. *Psychological Review*, *120*, 522–543. http://doi.org/10.1037/a0032457

Bhatia, S., & Mullett, T. L. (2016). Cognitive Psychology. *Cognitive Psychology*, *86*(C), 112–151. http://doi.org/10.1016/j.cogpsych.2016.02.002

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review, 113*, 700-765. http://doi.org/10.1037/0033-295X.113.4.700

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology*, *72*(4), 691.

Boring, E. G. (1921). The control of attitude in psychophysical experiments. *Psychological Review*, *27*, 440–452. http://dx.doi.org/10.1037/h0075451

Boring, E. G. (1921). The stimulus-error. *American Journal of Psychology*, *32*, 449–471. http://dx.doi.org/10.2307/1413768

Brewer, N., & Wells, G. L. (2011). Eyewitness identification. *Current Directions in Psychological Science*, *20*, 24-27.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153-178.

Brown, W. (1910). The judgement of difference. *University of California Publications in Psychology*, *1*, 1-71.

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, 32(2), 91-134.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432-459.

Clark, S. E. (1997). A familiarity-based account of confidence–accuracy inversions in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 232. http://dx.doi.org/10.1037/0278-7393.23.1.232

Clark, S. E. (2003). A memory and decision model for eyewitness identification. Applied Cognitive Psychology, 17, 629-654.

Clark, S. E., Howell, R. T., & Davey, S. L. (2008). Regularities in eyewitness identification. *Law and Human Behavior*, *32*(3), 187–218. http://doi.org/10.1007/s10979-006-9082-4

Dobbins, I. G., Kroll, N. E. A., & Liu, Q. (1998). Confidence–accuracy inversions in scene recognition: A remember–know analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24, 1306–1315.* http://dx.doi.org/10.1037/0278-7393.24.5.1306

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The over constraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16,* 1129–1135. z

Fernberger, S. W., 1930. (1930). The use of equality judgments in psychophysical procedures. *Psychological Review*, *37*, 107–112. http://dx.doi.org/10.1037/h0074662

Friedman, M. P., & Fleishman, E. A. (1956). A note on the use of a "don't know" alternative in multiple choice tests. *Journal of Educational Psychology, 47*, 344-349. http://doi.org/10.1037/h0043809

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., ... & Heiberger, R. (2012). Package 'car'. *Vienna: R Foundation for Statistical Computing*.

Garton, R., Reynolds, A., Hinder, M. R. & Heathcote, A. (2019). Equally flexible and optimal response bias in older compared to younger adults. *Psychology and Aging*, *34*, 821-835.

Hawkins, G.E. & Heathcote, A. (2019). Racing against the clock: Evidence-based vs. time-based decisions.

Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced choice episodic recognition. *Journal of Memory and Language*, *62*, 183-203. http://dx.doi.org/10.1016/j.jml.2009.11.003

Heathcote, A., Brown, S.D. & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In B. U. Forstmann, & E.-J. Wagenmakers (Eds). *An Introduction to Model-Based Cognitive Neuroscience (pp. 25-48).*New York, NY: Springer.

Heathcote, A., Lin, Y-S, Reynolds, A., Strickland, L., Gretton, M. & Matzke, D. (2018). Dynamic models of choice. *Behavior Research Methods*, *51*, 961-985. http://doi.org/10.3758/s13428-018-1067-y

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in psychology*, *3*, 292.

Higham, P. A. (2007). No Special K! A signal detection framework for the strategic regulation of memory accuracy. *Journal of Experimental Psychology: General*, *136*(1), 1–22. http://doi.org/10.1037/0096-3445.136.1.1

Horry, R., & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General*, *145*(12), 1615.

Jastrow, J. (1888). A critique of psycho-physic methods. *American Journal of Psychology*, *1*, 271-309. http://dx.doi.org/10.2307/1411321

Jessup, R. K., Veinott, E. S., Todd, P. M., & Busemeyer, J. R. (2009). Leaving the store empty-handed: Testing explanations for the too-much-choice effect using decision field theory. *Psychology and Marketing*, *26*(3), 299–320. http://doi.org/10.1002/mar.20274

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *Science*, *324*(5928), 759–764. http://doi.org/10.1126/science.1169405

Leite, F. P. & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. Judgment and Decision Making, 6, 651-687.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490-517.

Leite, F. P., & Ratcliff, R. (2011). What cognitive processes drive response biases? A diffusion model analysis. *Judgment & Decision Making*, *6*(7).

Malhotra, G., Leslie, D. S., Ludwig, C. J. H., & Bogacz, R. (2017). Time-varying decision boundaries: Insights from optimality analysis. *Psychonomic Bulletin and Review, 25*, 971-996. http://doi.org/10.3758/s13423-017-1340-6

Merkle, E. C., & Van Zandt, T. (2006). An application of the Poisson race model to confidence calibration. *Journal of Experimental Psychology: General*, *135*(3), 391.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, *78*, 99–147. http://doi.org/10.1016/j.cogpsych.2015.01.002

Osth, A. F., Bora, B., Dennis, S., & Heathcote, A. (2017). Diffusion vs. linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language*, *96*, 36-61.

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101-126.

Osth, A. F. & Farrell, S. (in press, 2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review*.

Palada, H., Neal, A., Vuckovic, A., Martin, R., Samuels, K. & Heathcote, A. (2016). Evidence accumulation in a complex task: Making choices about concurrent multi-attribute stimuli under time pressure. *Journal of Experimental Psychology: Applied*, *22*, 1-23. http://dx.doi.org/10.1037/xap0000074

Palada, H., Neal, A., Strayer, D., Ballard, T. & Heathcote, A. (accepted 3/May/2019). Competing for cognitive resources: Measuring workload in a time pressured dual-task environment. J*ournal of Experimental Psychology: Human Perception and Performance*.

Palada, H., Neal, A., Tay, R., & Heathcote, A. (2018). Understanding the causes of adapting, and failing to adapt, to time pressure in a complex multi-stimulus environment. *Journal of Experimental Psychology: Applied 24*, 380-399.. http://dx.doi.org/10.1037/xap0000176

Perfect, T. J., & Weber, N. (2012). How should witnesses regulate the accuracy of their identification decisions: One step forward, two steps back? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1810-1818. doi:10.1037/a0028461

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. http://doi.org/10.1037/a0019737

Provost, A. & Heathcote, A. (2015). Titrating decision processes in the mental rotation task, Psychological Review, 122, 735-754. http://doi.org/10.1037/a0039706

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1226–1243. http://doi.org/10.1037/a0036801

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356. http://doi.org/10.1111/1467-9280.00067

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological Review*, *116*, 59–83. http://doi.org/10.1037/a0014086

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination. *Psychological Review*, *120*, 697–719. http://doi.org/10.1037/a0033152

Shields, W. E., Smith, J. D., & Washburn, D. A. (1997). Uncertain responses by humans and Rhesus monkeys (Macaca mulatta) in a psychophysical same–different task. *Journal of Experimental Psychology: General*, *126*, 147–164. http://doi.org/10.1037/0096-3445.126.2.147

Shiffrin, R., Lee, M., Kim, W., & Wagenmakers, E.-J. (2008). A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods. *Cognitive Science*, *32*, 1248–1284. http://doi.org/10.1080/03640210802414826

Simen, P., Vlasov, K., & Papadakis, S. (2016). Scale (in)variance in a unified diffusion model of decision making and timing. *Psychological Review*, *123*, 151–181. http://doi.org/10.1037/rev0000014

Smith, J. D., Schull, J., Strote, J., McGee, K., Egnor, R., & Erb, L. (1995). The uncertain response in the bottlenosed dolphin (Tursiops truncatus). *Journal of Experimental Psychology: General; Journal of Experimental Psychology: General*, *124*, 391–408. http://doi.org/10.1037/0096-3445.124.4.391

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *B64*(4), 583–639. http://doi.org/10.1111/1467-9868.00353

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed-accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging, 25*, 377-390. http://doi.org/10.1037/a0018022

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal- variability and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64,* 1–34. doi:10.1016/j.cogpsych.2011.10.002

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior*, *20*, 479-496. http://dx.doi.org/10.1016/S0022-5371(81)90129-8

Tversky, A., & Shafir, E. (1993). Choice under conflict. *Psychological Science*, *3*, 358–361.

van Ravenzwaaij, D., Brown, S.D., Marley, A.J. & Heathcote, A. (in press). Accumulating advantages: A new approach to multialternative forced choice tasks, *Psychological Review*.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 582-600.

Van Zandt, T., & Maldonado-Molina, M. M. (2004). Response Reversals in Recognition Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30(6)*, 1147–1166. http://doi.org/10.1037/0278-7393.30.6.1147

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Vickers, D. (2001). Where does the balance of evidence lie with respect to confidence? In E. Sommerfeld, R. Kompass, & T. Lachmann (Eds.), *Proceedings of the 17th Annual Meeting of the International Society for Psychophysics* (pp. 148–153). Berlin: Pabst.

Watson, C. S., Kellogg, S. C., Kawanishi, D. T., & Lucas, P. A. (1973). The uncertain response in detection-oriented psychophysics. *Journal of Experimental Psychology*, *99*, 180–185. http://dx.doi.org/10.1037/h0034736

Weber, N., & Perfect, T. (2012). Improving eyewitness identification accuracy by screening out those who say they don't know. *Law and Human Behavior*, 36, 28-36. doi:10.1007/s10979-011-9269-1

Weber, N., & Perfect, T. J. (2013). Why telling a witness that it's OK to say they don't know is good for justice. *The Jury Expert: The Art and Science of Litigation and Advocacy*, *25*(3), 1-7.

White, C. M., Hoffrage, U., & Reisen, N. (2015). Choice deferral can arise from absolute evaluations or relative comparisons. *Journal of Experimental Psychology: Applied*, *21*(2), 140–157. http://doi.org/10.1037/xap0000043

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 385–398. http://doi.org/10.1037/a0034851

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262-276. doi: 10.1037/a0035940

Woodworth, R. S. (1937). Experimental Psychology. New York: Holt, 1938. *Department of Psychology Dartmouth College Hanover, New Hampshire*.